THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:
UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

August 16, 1999

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: ***60/093,940***
FILING DATE: ***July 23, 1998***

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS

P. SWAIN
Certifying Officer

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

Case No. GENSET.034PR
Date: July 23, 1998

ASSISTANT COMMISSIONER FOR PATENTS
WASHINGTON, D.C. 20231

ATTENTION: PROVISIONAL PATENT APPLICATION

Sir:

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR § 1.53(b)(2).

For: **A NUCLEIC ACID ENCODING A GERANYL-GERANYL PYROPHOSPHATE SYNTHETASE (GGPPS) AND POLYMORPHIC MARKERS ASSOCIATED WITH SAID NUCLEIC ACID**

Name of Sole Inventor: Lydie Bougueleret
Residence Address: 108, avenue Victor Hugo, 92170 Vanves, France

Enclosed are:

(X) Specification in 62 pages.

(X) Sequence Listing in 33 pages.

(X) Sequence Submission in 1 page.

(X) Sequence Listing in computer readable form.

(X) Two (2) sheets of drawings.

(X) A check in the amount of $150 to cover the filing fee is enclosed.

(X) A return prepaid postcard.

(X) The Commissioner is hereby authorized to charge any additional fees which may be required, or credit any overpayment to Account No. 11-1410. A duplicate copy of this sheet is enclosed.

Was this invention made by an agency of the United States Government or under a contract with an agency of the United States Government?

(X) No.

( ) Yes. The name of the U.S. Government agency and the Government contract number are:

KNOBBE, MARTENS, OLSON & BEAR, LLP
620 NEWPORT CENTER DR 16TH FLOOR NEWPORT BEACH, CA 92660
(949) 760-0404 FAX (949) 760-9502

(X) Please send correspondence to:

Daniel Hart
Knobbe, Martens, Olson & Bear, LLP
620 Newport Center Dr., 16th Floor
Newport Beach, CA 92660

Respectfully submitted,

Daniel Hart

Daniel Hart
Registration No. 40,637

S:\DOCS\DOH\DOH-1896.DOC
072398

A NUCLEIC ACID ENCODING A GERANYL-GERANYL PYROPHOSPHATE SYNTHETASE (GGPPS) AND POLYMORPHIC MARKERS ASSOCIATED WITH SAID NUCLEIC ACID.

FIELD OF THE INVENTION

The present invention relates to a purified or isolated polynucleotide encoding human geranylgeranyl pyrophosphate synthetase, the regulatory nucleic acids contained therein, a polymorphic marker thereof and the resulting encoded protein, as well as to methods and kits for detecting this polynucleotide and this protein. The present invention also pertains to a polynucleotide carrying the natural regulatory regions of the *hGGPS* gene which is useful, for example, to express a heterologous nucleic acid in host cells or host organisms as well as functionally active regulatory polynucleotides derived from said regulatory region. The invention also consists in genetic markers, namely biallelic markers, which may be useful for the diagnosis of diseases related to an alteration in the regulatory or coding regions of *hGGPS*, such as pathologies related to a defect in the mevalonic biosynthetic pathway.

Throughout this application, various references are referred to within parentheses. The disclosures of these publications in their entireties are hereby incorporated by reference into this application to more fully describe the sate of the art to which this invention pertains.

BACKGROUND OF THE INVENTION

Prenylation is the least common known lipid modification. Other lipid modifications include palmitylation, myristylation and glycophospholipidation. However, prenylation is a surprisingly common form of post-translational protein modification with an occurrence of 0.5 % of all cellular proteins. Prenylation is a covalent modification which involves the attachment of either a C15 farnesyl or a C20 geranylgeranyl isoprenoid, both being products of the mevalonic acid biosynthetic pathway, to one or more cysteine residues at the carboxyl terminus of the protein via a thioether bond. The C20 geranylgeranyl modification predominates over the C15 farnesyl modification in terms of frequency of occurrence. The structural environment of the cysteine residue determines the specific type and number of isoprenoid groups that attach to each

cysteine. The covalent modification resulting from prenylation renders proteins more hydrophobic and, together with a subsequent modification cascade, facilitates their association with membranes. Protein prenylation also mediates protein-protein interactions. Prenylated proteins can be involv d in signal transduction, intracellular vesicular transport, cytoskeletal organization, cell growth control and polarity, viral replication and protein folding/assembly. In mammals, prenylated proteins are more frequently modified by one or more geranylgeranyl groups. Farnesylation has only been found to occur in the retinal heterotrimeric G protein transducin, in retinal rhodopsin kinase, in *ras* proteins, in nuclear lamins, and in yeast mating factors. Geranylgeranylation is found in all of the remaining heterotrimeric G proteins and small G proteins.

Heterotrimeric G-proteins which are required for intracellular signal transduction between receptors and effector enzymes present one or two prenylated subunits. This modification is often required for association of the functional complex with the membrane.

Among small G proteins, *Ras* proteins, which comprise oncogenic forms, regulate signal transduction pathways controlling cell proliferation and differentiation. All *ras* proteins are prenylated and this modification is critical for their transport to the inner surface of the plasma membrane and their biological functions.

Other prenylated proteins belonging to the *ras* protein superfamily are involved in the regulation of intracellular vesicular transport (Rab/YPT1), in the cytoskeletal organization of polymerized actin to produce stress fibers (Rho) or membrane ruffling (Rac), in the oxydative burst of phagocytic cells (Rac), in the control of the cell cycle and polarity (cdc24Hs/G25K), and in negative growth control (Rap/Krev-1). Prenylation is important to these activities. For example, Rab/YPT prenylation is critical for the association of these proteins with specific intracellular compartments and in their regulation of intracellular transport processes.

One hypothesis is that rather than providing only an increase in hydrophobicity, the isoprenoid acts as part of a recognition unit for specific receptors that interact with either farnesylated or geranylgeranylated proteins. The recent observations that geranylgeranyl-modified forms of K-Ras4B or H-Ras proteins exhibit intracellular localizations which are different from those of their authentic farnesylated counterparts is consistent with this possibility.

Moreover, prenylation of nuclear lamins, which are involved in the mitotic control of membrane assembly, is necessary for the proper assembly of these proteins

into the nuclear lamina. Indeed, prenylation is necessary to the maturation by cleavage of prelamin A in lamin A and to obtain functional lamin B.

Geranylgeranyl pyrophosphate synthetase (GGPS) is involved in the mevalonic acid biosynthetic pathway and is located in the cytosol. It catalyzes the consecutive condensation of isopentenyl diphosphate with allylic diphosphates to produce GGPP. This biosynthesis of GGPPS is regulated according to requirements for protein prenylation. GGPS has been found to be expressed in human fetal heart, as described in the PCT Application No WO 96/21736.

SUMMARY OF THE INVENTION

The invention concerns a nucleic acid molecule comprising the genomic sequence of a human geranylgeranyl pyrophosphate synthetase gene.

hGGPS gene, corresponding cDNAs and regulatory nucleotide sequences.

As shown in Figure 1, the *hGGPS* genomic sequence comprises a regulatory sequence preceding the ORF encoding the hGGPS protein and another regulatory sequence localized downstream of the *hGGPS* ORF.

The present invention first concerns a purified or isolated nucleic acid comprising a nucleotide sequence of SEQ ID No 1, or a nucleotide sequence complementary thereto. The *hGGPS* genomic sequence is depicted in the upper line of Figures 1 and 2. The transcription of this genomic sequence leads to more than one mRNA final product, due to alternative splicing events, as it is described below..

In Figure 1, four exons are represented in the upper line as vertical bars, namely Exon 1, Exon 2, Exon 3 and Exon 4. These four exons are those contained in a first *hGGPS* mRNA molecule detected by the inventors (see middle line of Figure 1), and more precisely in the mRNA molecule of the nucleotide sequence of SEQ ID No 4.

In Figure 2, four exons are represented in the upper line as vertical bars, respectively Exon 1*bis*, Exon 2, Exon 3 and Exon 4. These four exons are those included in a second *hGGPS* mRNA molecule detected by the inventors (see middle line of Figure 2), and more precisely in the mRNA molecule of the nucleotide sequence of SEQ ID No 5.

Consequently, another object of the invention consists in a purified or isolated nucleic acid comprising a nucleotide sequence selected from the group consisting of SEQ ID Nos 4 and 5.

Another object of the invention consists in a purified or isolated nucleic acid comprising a nucleic acid fragment of a nucleotide sequence selected from the group consisting of SEQ ID Nos 4 and 5, wherein this nucleic acid fragment encodes a polypeptide having an amino acid sequence beginning at the amino acid in position 200 and ending at the amino acid in position 300 of the hGGPS polypeptide of SEQ ID No 6, or a nucleic acid encoding a peptide fragment thereof.

The invention further deals with a regulatory nucleic acid comprising a nucleotide sequence flanking the ORF sequence contained in the *hGGPS* gene of SEQ ID No 1. The invention thus encompasses a purified or isolated nucleic acid comprising a regulatory polynucleotide which is selected from the group consisting of the nucleotide sequences of SEQ ID Nos 2 and 3.

The present invention is also directed to a polynucleotide comprising a functional portion of a regulatory region contained in the contemplated *hGGPS* gene and to its use in a recombinant expression vector carrying a polynucleotide encoding a polypeptide or a nucleic acid of interest.

A further object of the invention consists in polynucleotide fragments of the *hGGPS* gene, preferably polynucleotide fragments located outside the *hGGPS* ORF, that are useful for detecting the presence of an unaltered or altered copy of this gene within the human genome of a given individual and also for the detection and/or quantification of the expression of *hGGPS* in said individual host organism.

When used herein, an altered copy of the *hGGPS* gene according to the invention is intended to designate the *hGGPS* gene that has undergone at least one substitution or deletion of one or several nucleotides, wherein said nucleotide substitution, addition or deletion of one or several nucleotides causes a change in the amino acid sequence of SEQ ID No 6 or alternatively causes an increase or a decrease in the expression of the *hGGPS* gene.

Biallelic markers

The invention also relates to a nucleotide sequence, preferably a purified and/or isolated polynucleotide comprising a sequence defining a biallelic marker located in the sequence of the *hGGPS* gene, a fragment or variant thereof or a sequence complementary thereto. As used herein, the terminology "defining a biallelic marker" means that a sequence includes a polymorphic base from a biallelic marker. The sequences defining a biallelic marker may be of any length consistent with their intended use, provided that they contain a polymorphic base from a biallelic mark r.

The sequence has between 1 and 500 nucleotides in length, preferably between 5; 10, 15, 20, 25 or 40 and 200 nucleotides and more preferably between 30 and 50 nucleotides in length. Preferably, the sequences defining a biallelic marker include the polymorphic base of one of SEQ ID Nos 7-8. In some embodiments the sequences defining a biallelic marker comprise one of the sequences selected from the group consisting of SEQ ID Nos 7-8. Likewise, the term "marker" or "biallelic marker" requires that the sequence is of sufficient length to practically (although not necessarily unambiguously) identify the polymorphic allele, which usually implies a length of at least 4, 5, 6, 10, 15, 20, 25 or 40 nucleotides.

The invention further concerns a nucleic acid encoding a hGGPS protein, wherein said nucleic acid comprises a nucleotide sequence selected from the group consisting of SEQ ID Nos 7-8.

The invention also relates to nucleotide sequence selected from the group consisting of SEQ ID Nos 7-8 or a fragment or a variant thereof.

The invention also pertains to a nucleotide sequence selected from the group consisting of a variant or fragment of SEQ ID Nos 7-8, said fragment comprising at least 8 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID Nos 7-8 and including the polymorphic base thereof.

Identification and characterization of further biallelic markers

Another aspect of the present invention is a method of identifying biallelic markers in the genomic region harboring the *hGGPS* gene comprising the steps of :

- designing a plurality of primer sequences capable of amplifying portions of the genomic region containing the *hGGPS* gene, and in particular portions of the polynucleotide of SEQ ID No 1;
- amplifying portions of the genomic region containing the *hGGPS* gene from a plurality of individuals using said primers to obtain a plurality of amplicons; and
- sequencing said plurality of amplicons to identify biallelic markers in the genomic region harboring the *hGGPS* gene.

Oligonucleotide probes and primers

The invention also relates to oligonucleotide molecules useful as probes or primers, wherein said oligonucleotide molecules hybridize specifically with a nucleotide sequence selected from the group consisting of the regulatory polynucleotides of the

invention, said group including the nucleotide sequences of SEQ ID Nos 2 and 3 and their fragments and variants.

More precisely, a nucleic acid probe according to the invention comprises at least 8 consecutive nucleotides of a regulatory polynucleotide as defined above, preferably from 8 to 200 consecutive nucleotides, more particularly from 10, 15, 20 or 30 to 100 consecutive nucleotides, more preferably from 10 to 50 nucleotides, and most preferably from 15 to 30 consecutive nucleotides of a regulatory polynucleotide according to the present invention.

The invention further concerns detection or amplification kits containing a pair of oligonucleotide primers or an oligonucleotide probe according to the invention. The kits of the present invention can also comprise optional elements including appropriate amplification reagents such as DNA polymerases when the kit comprises primers, or reagents useful in hybridization between a labeled hybridization probe and the *hGGPS* gene

Amplification of a polynucleotide of the invention

The invention also concerns a method for the amplification of a regulatory or a coding region of the *hGGPS* gene or a fragment or a variant thereof in a test sample. The method comprises the steps of :

- contacting a test sample suspected of containing the desired *hGGPS* sequence or portion thereof with amplification reaction reagents comprising a pair of amplification primers such as those described above, the primers being located on either side of the *hGGPS* nucleotide region to be amplified. The method may further comprise the step of detecting the amplification product. For example, the amplification product may be detected using a detection probe that can hybridize with an internal region of the amplicon sequences. Alternatively, the amplification product may be detected with any of the primers used for the amplification reaction themselves, optionally under a labeled form.

Suitable primers include the nucleic acids of SEQ ID Nos 9-11, these primers being located on either side of a biallelic marker according to the invention. The method may further comprise the step of detecting the amplification product. For example, the amplification product may be detected using a detection probe that can hybridize with an internal region of the amplicon sequences.

Vectors and host cells

A further object of the present invention is a recombinant expression vector for the expression of an heterologous polynucleotide, wherein said vector comprises a nucleic acid comprising a nucleotide sequence of SEQ ID No 2, or biologically active nucleotide fragments and variants thereof. The heterologous polynucleotide codes either for a desired polypeptide of interest or for a polynucleotide, for example a sense or an antisense DNA molecule. Optionally, such a recombinant expression vector may also comprise a nucleic acid comprising a nucleotide sequence of SEQ ID Nos 3 and biologically active nucleotide fragments and variants thereof.

The invention further deals with a recombinant expression vector for the expression of a nucleotide sequence comprising a polynucleotide of SEQ ID No 3 or a biologically active fragment or variant thereof.

Another recombinant vector according to the invention consists in a recombinant expression vector that comprises a nucleic acid comprising a polynucleotide selected from the group consisting of the nucleotide sequences of SEQ ID Nos 4 and 5.

A further recombinant vector according to the invention comprises a purified or isolated nucleic acid comprising a nucleic acid fragment of a nucleotide sequence selected from the group consisting of SEQ ID Nos 4 and 5, wherein this nucleic acid fragment encodes a polypeptide having an amino acid sequence beginning at the amino acid in position 200 and ending at the amino acid in position 300 of the hGGPS polypeptide of SEQ ID No 6, or a nucleic acid encoding a peptide fragment thereof.

hGGP polypeptide of the invention

The invention also concerns a purified or isolated hGGPS polypeptide encoded by a nucleic acid selected from the group consisting of SEQ ID Nos 4 and 5.

More particularly, the invention relates to a purified or isolated hGGPS polypeptide consisting of the aminoacid sequence of SEQ ID No 6. This polypeptide differs from the hGGPS described in the PCT Patent Application No WO 96/21736 mainly in its C-terminal portion and particularly in the C-terminal portion beginning at the amino acid in position 200 and ending at the amino acid in position 300 of the hGGPS polypeptide of SEQ ID No 6.

As used herein, the term "isolated" requires that the material be removed from its original environment (e.g. the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal

is not isolated, but the same polynucleotide or DNA or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotide could be part of a vector and/or such polynucleotide or polypeptide could be part of a composition and still be isolated in that the vector or composition is not part of its natural environment.

Throughout the present specification, the expression "nucleotide sequence" may be employed to designate indifferently a polynucleotide or an nucleic acid. More precisely, the expression "nucleotide sequence" encompasses the nucleic material itself and is thus not restricted to the sequence information (i.e. the succession of letters chosen among the four base letters) that biochemically characterizes a specific DNA or RNA molecule.

Antibodies

The invention also concerns a purified or isolated antibody which is capable of specifically binding to the GGPS protein comprising the amino acid sequence of SEQ ID No 6 or which is capable of specifically binding to a C-terminal fragment of said protein, and more particularly to a peptide fragment comprised in the polypeptide beginning at the amino acid in position 200 and ending at the amino acid in position 300 of the amino acid sequence of SEQ ID No 6.

The invention also deals with methods and kits for detecting the presence of the polypeptide comprising the amino acid sequence SEQ ID No 6 in a test sample.

The method particularly comprises contacting a test sample suspected of containing the amino acid sequence of SEQ ID No 6 with an antibody of the invention.

The kit comprises an antibody of the invention and preferably means for revealing the formation of an antigen-antibody complex.

Complementary polynucleotides

For the purpose of the present invention, a first polynucleotide is deemed to be complementary to a second polynucleotide when each base in the first polynucleotide is paired with its complementary base. Complementary bases are, generally, A and T (or A and U), or C and G.

Methods for screening substances or molecules modulating the expression f hGGPS.

Another object of the present invention consists of methods and kits for the screening of candidate substances that are able to modulate the expression of *hGGPS*.

The present invention also concerns a method for screening substances or molecules that are able to increase, or in contrast to decrease or even suppress the expression of the *hGGPS* gene. Such a method may allow the one skilled in the art to select substances exerting a positive or negative regulating effect on the expression level of the *hGGPS* gene and thus enabling a correction in the *hGGPS* expression levels in individuals in which the *hGGPS* expression is defective (i.e. lower or in contrast higher than the normal expression levels).

Thus, is also part of the present invention a method for screening of a candidate substance or molecule that modulates the expression of the *hGGPS* gene according to the invention, wherein this method comprises the following steps:

a) providing a recombinant host cell containing a nucleic acid, wherein said nucleic acid comprises a nucleotide sequence of SEQ ID No 2 or a biologically active fragment or variant thereof operably linked to a polynucleotide encoding a detectable protein;

b) obtaining a candidate substance, and

c) determining the ability of the candidate substance to modulate the expression levels of the polynucleotide encoding the detectable protein.

Among the preferred polynucleotides encoding a detectable protein, there may be cited polynucleotides encoding beta galactosidase, green fluorescent protein (GFP) and chloramphenicol acetyl transferase (CAT).

Therefore, the invention also pertains to a kit for the screening of a candidate substance or molecule modulating the expression of the *hGGPS* gene, wherein said kit comprises a recombinant vector containing a polynucleotide encoding a detectable protein under the control of a nucleotide sequence of SEQ ID No 2 or a biologically active fragment or variant thereof.

Preferably, the regulatory sequence contained in the recombinant vector described above is located upstream the polynucleotide encoding a detectable protein.

Another embodiment of a method for screening candidate substances or molecules modulating the expression of the *hGGPS* gene comprises the following steps :

a) providing a recombinant host cell expressing a nucleic acid, wherein said nucleic acid comprises a nucleotide sequence selected from the group consisting of SEQ ID Nos 1, 4 and 5;

b) obtaining a candidate substance, and

c) determining the ability of the candidate substance to modulate the expression levels of the nucleotide sequence selected from the group consisting of SEQ ID Nos 1, 4 and 5.

The invention also deals with a kit for the screening of a candidate substance or molecule modulating the expression of the *hGGPS* gene, wherein said kit comprises a recombinant vector that allows the expression of a nucleotide sequence selected from the group consisting of SEQ ID Nos : 1, 4 and 5 or alternatively a recombinant host cell containing such a recombinant vector.

For the design of suitable recombinant vectors useful for performing the screening methods described above, it will be referred to the section of the present specification wherein the preferred recombinant vectors of the invention are described in more detail.

Variants and fragments of the polynucleotides according to the invention.

The invention also relates to variants and fragments of the polynucleotides described herein.

Variants of polynucleotides, as the term is used herein, are polynucleotides that differ from a reference polynucleotide. A variant of a polynucleotide may be a naturally occurring variant such as a naturally occurring allelic variant, or it may be a variant that is not known to occur naturally. Such non-naturally occurring variants of the polynucleotide may be made by mutagenesis techniques, including those applied to polynucleotides, cells or organisms. Generally, differences are limited so that the nucleotide sequences of the reference and the variant are closely similar overall and, in many regions, identical.

Variants of polynucleotides according to the invention include, without being limited to, nucleotide sequences at least 95% identical to a nucleic acid selected from the group consisting of SEQ ID Nos 1-5 or to any polynucleotide fragment of at least 8 consecutive nucleotides from a nucleic acid selected from the group consisting of SEQ ID Nos 2 and 3, and preferably at least 99% identical, more particularly at least 99.5% identical, and most preferably at least 99.8% identical to a nucleic acid selected from

the group consisting of SEQ ID Nos 2 and 3 or to any polynucleotide fragment of at least 8 consecutive nucleotides of these nucleic acids.

A polynucleotide fragment is a polynucleotide having a sequence that entirely is the same as part but not all of a given nucleotide sequence, preferably the nucleotide sequence of a nucleic acid selected from the group consisting of SEQ ID Nos 2 and 3. The fragment is preferably a portion of the regulatory sequences of the *hGGPS* gene.

Such fragments may be "free-standing", i.e. not part of or fused to other polynucleotides, or they may be comprised within a single larger polynucleotide of which they form a part or region. However, several fragments may be comprised within a single larger polynucleotide.

As representative examples of polynucleotide fragments of the invention, there may be mentioned those which have from about 4, 6, 8, 15, 20, 25, 40, 10 to 20, 10 to 30, 30 to 55, 50 to 100, 75 to 100 or 100 to 200 nucleotides in length.

BRIEF DESCRIPTION OF THE DRAWING

Figure 1 : Map of the genomic, cDNA and coding (CDS) sequences of *hGGPS* : (1) upper line, genomic sequence; (2) cDNA sequence of SEQ ID No 4; (3) coding sequence (CDS).

Figure 2 : Map of the genomic, cDNA and coding (CDS) sequences of *hGGPS* : (1) upper line, genomic sequence; (2) cDNA sequence of SEQ ID No 5; (3) coding sequence (CDS).

DETAILED DESCRIPTION OF THE INVENTION

The *hGGPS* gene of the invention is located on chromosome 1, and mor precisely on the 1q42-1q43 locus of this chromosome. This chromosome 1 locus has been shown to carry a predisposing gene for prostate cancer (Berthon et al., 1998).

The *hGGPS* gene of the invention is located in the vicinity of a retinoblastoma binding protein gene. Indeed, the coding sequence of this latter gene is on a strand which is opposite to the strand carrying the *hGGPS* Open Reading Frame.

The aim of the present invention is to provide polynucleotides derived from the *hGGPS* gene, particularly those useful to design suitable means for detecting the presence of this gene in a test sample or alternatively to discriminate between the *hGGPS* mRNA molecules that are present in a test sample. Other polynucleotides of the invention are useful to design suitable means to express a desired polynucleotide

of interest. The invention also relates to the hGGPS polypeptide having the amino acid sequence of SEQ ID No 6.

hGGPS gene polynucleotide, cDNAs and associated regulatory regions.

Genomic sequences

The invention concerns a purified or isolated nucleic acid encoding the hGGPS polypeptide, wherein said nucleic acid comprises the nucleotide sequence of SEQ ID No 1.

The invention also encompasses a purified or isolated nucleic acid having at least 95% nucleotide identity with the nucleotide sequence of SEQ ID No 1. The nucleotide differences as regards to the nucleotide sequence of SEQ ID No 1 are generally randomly distributed throughout the entire nucleic acid. Nevertheless, preferred nucleic acids are those wherein the nucleotide differences as regards to the nucleotide sequence of SEQ ID No 1 are predominantly located outside the coding sequences contained in Exons 2, 3 and 4.

As already mentioned, the *hGGPS* genomic nucleic acid sequence comprises five exons. Exon 1 starts at the nucleotide in position 486 and ends at the nucleotide in position 546 of the nucleotide sequence of SEQ ID No1; Exon 1*bis* starts at the nucleotide in position 633 and ends at the nucleotide in position 826 of the nucleotide sequence of SEQ ID No 1; Exon 2 starts at the nucleotide in position 7292 and ends at the nucleotide in position 7384 of the nucleotide sequence of SEQ ID No1; Exon 3 starts at the nucleotide in position 13760 and ends at the nucleotide in position 13830 of the nucleotide sequence of SEQ ID No 1; Exon 4 starts at the nucleotide in position 14063 and ends at the nucleotide in position 15251 of the nucleotide sequence of SEQ ID No1.

The hGGPS introns defined hereinafter for the purpose of the present invention are not exactly what is generally understood as "introns" by the one skilled in the art and will consequently be defined below.

Generally, an intron is defined as a nucleotide sequence that is present both in the genomic DNA and in the unspliced mRNA molecule, and which is absent from the mRNA molecule which has undergone the splicing events. In the case of the hGGPS gene, the inventors have found that at least two different spliced mRNA molecules are produced when this gene is transcribed, as it will be described in detail in a further

section of the specification. The first spliced mRNA molecule comprises Exons 1, 2, 3 and 4, as shown in Figure 1. Thus, the genomic nucleotide sequence comprised between Exon 1 and Exon 2 is an intronic sequence as regards to this first mRNA molecule, despite the fact that this intronic sequence contains Exon 1*bis*. In contrast, Exon 1*bis* is of course an exonic nucleotide sequence as regards to the second hGGPS mRNA molecule shown in Figure 2.

For the purpose of the present invention and in order to make a clear and unique designation of the different nucleic acids of the invention, it has been postulated that the polynucleotides contained both in the nucleotide sequence of SEQ ID No 1 and in any of the nucleotide sequences of SEQ ID Nos 4 or 5 are considered as exonic sequences. Conversely, the polynucleotides contained in the nucleotide sequence of SEQ ID No. 1 and located between Exon 1 and Exon 4, but which are absent both from the nucleotide sequence of SEQ ID No 4 and from the nucleotide sequence of SEQ ID No 5 are considered as intronic sequences.

Consequently, Intron 1 (nucleotide sequence located between Exon 1 and Exon 1bis) starts at the nucleotide in position 547 and ends at the nucleotide in position 632 of the nucleotide sequence of SEQ ID No 1; Intron 1*bis* starts at the nucleotide in position 827 and ends at the nucleotide in position 7291 of the nucleotide sequence of SEQ ID No 1. Intron 2 starts at the nucleotide in position 7385 and ends at the nucleotide in position 13761 of the nucleotide sequence of SEQ ID No 1. Intron 3 starts at the nucleotide in position 13831 and ends at the nucleotide in position 14064 of the nucleotide sequence of SEQ ID No 1.

The nucleic acids defining the *hGGPS* introns described above, as well as their fragments and variants, may be used as oligonucleotide primers or probes in order to detect the presence of a copy of the *hGGPS* in a test sample, or alternatively in order to amplify a target nucleotide sequence within the *hGGPS* intronic sequences.

***hGGPS* cDNAs**

The inventors have discovered that the expression of the *hGGPS* gene leads to the production of at least two mRNA molecules, respectively a first and a second *hGGPS* transcription product.

The first transcription product comprises Exons 1, 2, 3 and 4. This cDNA of SEQ ID No 4 includes a 5'-UTR region, spanning the whole Exon 1 and part of Exon 2.

This 5'-UTR region starts from the nucleotide at position 1 and ends at the nucleotide in position 84 of SEQ ID No 4. The cDNA of SEQ ID No 4 includes a 3'-UTR region starting from the nucleotide at position 988 and ending at the nucleotide at position 1414 of SEQ ID No 4. The ORF encoding hGGPS is comprised between the nucleotide in position 85 and the nucleotide in position 987 of SEQ ID No 4.

The second transcription product comprises Exons *1bis*, 2, 3 and 4. This cDNA of SEQ ID No 5 includes a 5'-UTR region starting from the nucleotide at position 1 and ending at the nucleotide in position 217 of SEQ ID No 5. The cDNA of SEQ ID No 6 includes a 3'-UTR region starting from the nucleotide at position 1121 and ending at the nucleotide at position 1547 of SEQ ID No 5. The ORF encoding hGGPS is comprised between the nucleotide in position 218 and the nucleotide in position 1120 of the nucleotide sequence of SEQ ID No 5.

Another object of the invention consists of a purified or isolated nucleic acid selected from the group consisting of the nucleotide sequences of SEQ ID Nos 4 and 5 or non coding fragments thereof.

The invention also pertains to a purified or isolated nucleic acid having at least 95% of nucleotide identity with any of the nucleotide sequences of SEQ ID Nos 4 and 5.

The nucleotide differences as regards to the nucleotide sequences of SEQ ID Nos 4 and 5 are generally randomly distributed throughout the entire nucleic acid. Nevertheless, preferred nucleic acids are those wherein the nucleotide differences as regards to the nucleotide sequence of SEQ ID No 1 are predominantly located outside the coding sequences, and more precisely in the 5'-UTR and the 3'-UTR sequences contained in either nucleotide sequences of SEQ ID Nos 4 and 5.

Regulatory sequences

As already mentioned hereinbefore, the polynucleotide of SEQ ID No 1 contains regulatory sequences both in the non-coding 5'-flanking region and in the non-coding 3'-flanking region that border the *hGGPS* coding region.

The longest 5'-regulatory sequence of the *hGGPS* gene comprises the nucleotide sequence of SEQ ID No 2. The polynucleotide sequence of SEQ ID No 2 is localized between the nucleotide in position 1 and the nucleotide in position 7314 of SEQ ID No1. This polynucleotide sequence contains the transcription and the

translation start sites as well as the 5'-UTR region of the two identified cDNAs of SEQ ID Nos 4 and 5.

The *hGGPS* 3'-regulatory region, as shown in Figur 1, comprises a nucleotide sequence starting from the nucleotide in position 14825 of SEQ ID No 1 and ending at the nucleotide in position 17131 of SEQ ID No1, such nucleotide sequence consisting in the nucleic acid of SEQ ID No 3.

Such a *hGGPS* 3'-regulatory region defined above comprises the 3'-UTR region which is common to the cDNAs of SEQ ID Nos 4 and 5.

Polynucleotides derived from the *hGGPS* regulatory regions described above are useful in order to detect the presence of at least a copy of any of the nucleotide sequences of SEQ ID Nos 1, 4 or 5 in a test sample.

Thus, a further object of the invention consists in a purified or isolated nucleic acid of at least eight nucleotides in length, wherein said nucleic acid hybridizes under stringent hybridization conditions with a polynucleotide selected from the group consisting of the nucleotide sequences of SEQ ID Nos 2 and 3, or a sequence complementary thereto.

For the purpose of defining such a hybridizing nucleic acid according to the invention, the stringent hybridization conditions are the followings :

the hybridization step is realized at 65°C in the presence of 6 x SSC buffer, 5 x Denhardt's solution, 0,5% SDS and 100µg/ml of salmon sperm DNA.

The hybridization step is followed by four washing steps :

- two washings during 5 min, preferably at 65°C in a 2 x SSC and 0.1%SDS buffer;
- one washing during 30 min, preferably at 65°C in a 2 x SSC and 0.1% SDS buffer;
- one washing during 10 min, preferably at 65°C in a 0.1 x SSC and 0.1%SDS buffer.

These hybridization conditions are suitable for a nucleic acid molecule of about 20 nucleotides in length. There is no need to say that the hybridization conditions described above are to be adapted according to the length of the desired nucleic acid, following techniques well known to the one skilled in the art. The suitable hybridization conditions may for example be adapted according to the teachings disclosed in the book of Hames and Higgins (1985).

The promoter activity of the regulatory regions contained in the *hGGPS* nucleotide sequence of SEQ ID No 1 can be assessed as described below.

Genomic sequences located upstream of the *hGGPS* gene are cloned into a suitable promoter reporter vector, such as the pSEAP-Basic, pSEAP-Enhancer, pβgal-Basic, pβgal-Enhancer, or pEGFP-1 Promoter Reporter vectors available from Clontech. Briefly, each of these promoter reporter vectors include multiple cloning sites positioned upstream of a reporter gene encoding a readily assayable protein such as secreted alkaline phosphatase, beta galactosidase, or green fluorescent protein. The sequences upstream the *hGGPS* coding region are inserted into the cloning sites upstream of the reporter gene in both orientations and introduced into an appropriate host cell. The level of reporter protein is assayed and compared to the level obtained from a vector which lacks an insert in the cloning site. The presence of an elevated expression level in the vector containing the insert with respect to the control vector indicates the presence of a promoter in the insert. If necessary, the upstream sequences can be cloned into vectors which contain an enhancer for increasing transcription levels from weak promoter sequences. A significant level of expression above that observed with the vector lacking an insert indicates that a promoter sequence is present in the inserted upstream sequence.

Promoter sequences within the upstream genomic DNA may be further defined by constructing nested deletions in the upstream DNA using conventional techniques such as Exonuclease III digestion. The resulting deletion fragments can be inserted into the promoter reporter vector to determine whether the deletion has reduced or obliterated promoter activity. In this way, the boundaries of the promoters may be defined. If desired, potential individual regulatory sites within the promoter may be identified using site directed mutagenesis or linker scanning to obliterate potential transcription factor binding sites within the promoter individually or in combination. The effects of these mutations on transcription levels may be determined by inserting the mutations into cloning sites in promoter reporter vectors.

Polynucleotides carrying the regulatory elements located both at the 5' end and at the 3' end of the *hGGPS* coding region may be advantageously used to control the transcriptional and translational activity of an heterologous polynucleotide of interest.

Thus, the present invention also concerns a purified or isolated nucleic acid comprising a polynucleotide which is selected from the group consisting of the nucleotide sequences SEQ ID Nos 2 and 3, or a sequence complementary thereto or a biologically active fragment or variant thereof.

Preferred fragments of the nucleic acid of SEQ ID No 2 have a length of about 400 nucleotides, more particularly about 300 nucleotides, more preferably 200 nucleotides and most preferably about 100 nucleotides.

Preferred fragments of the nucleic acid of SEQ ID No 3 have a length of about 600 nucleotides, more particularly about 300 nucleotides, more preferably 200 nucleotides and most preferably about 100 nucleotides.

By a "biologically active fragment" of SEQ ID Nos 2 and 3 according to the present invention is intended a polynucleotide comprising or alternatively consisting in a fragment of said polynucleotide which is functional as a regulatory region for expressing a recombinant polypeptide or a recombinant polynucleotide in a recombinant cell host.

For the purpose of the invention, a nucleic acid or polynucleotide is "functional" as a regulatory region for expressing a recombinant polypeptide or a recombinant polynucleotide if said regulatory polynucleotide contains nucleotide sequences which contain transcriptional and translational regulatory information, and such sequences are "operably linked" to nucleotide sequences which encode the desired polypeptide or the desired polynucleotide. An operable linkage is a linkage in which the regulatory nucleic acid and the DNA sequence sought to be expressed are linked in such a way as to permit gene expression.

More precisely, two DNA molecules (such as a polynucleotide containing a promoter region and a polynucleotide encoding a desired polypeptide or polynucleotide) are said to be "operably linked" if the nature of the linkage between the two polynucleotides does not (1) result in the introduction of a frame-shift mutation or (2) interfere with the ability of the polynucleotide containing the promoter to direct the transcription of the coding polynucleotide. The promoter polynucleotide would be operably linked to a polynucleotide encoding a desired polypeptide or a desired polynucleotide if the promoter is capable of effecting transcription of the polynucleotide of interest.

In order, to identify the relevant biologically active polynucleotide derivatives of SEQ ID Nos 2 and 3, the one skill in the art will refer to the book of Sambrook et al. (Sambrook, J. Fritsch, E. F., and T. Maniatis. 1989. Molecular cloning: a laboratory manual. 2ed. Cold Spring Harbor Laboratory, Cold spring Harbor, New York) which describes the use of a recombinant vector carrying a marker gene (i.e. beta galactosidase, chloramphenicol acetyl transferase, etc.) the expression of which will be

detected when placed under the control of a biologically active derivative polynucleotide of SEQ ID Nos 2 and 3.

The regulatory polynucleotides of the invention may be prepared from any of the nucleotide sequence SEQ ID Nos 1-3 or one of the nucleotide sequences SEQ ID Nos 4 and 5 by cleavage using suitable restriction enzymes, as described for example in the book of Sambrook et al. (1989).

The regulatory polynucleotides may also be prepared by digestion of any of the SEQ ID Nos 1-3 or SEQ ID Nos 4 and 5 by an exonuclease enzyme, such as for example Bal31 (Wabiko et al., 1986, DNA, 5(4):305-314).

These regulatory polynucleotides can also be prepared by nucleic acid chemical synthesis, as described elsewhere in the specification, where oligonucleotide probes or primers synthesis is disclosed.

The regulatory polynucleotides according to the invention may be advantageously part of a recombinant expression vector that may be used to express a coding sequence in a desired host cell or host organism. The recombinant expression vectors according to the invention are described elsewhere in the specification.

A preferred 5'-regulatory polynucleotide of the invention includes the 5'-untranslated region (5'-UTR) located between the nucleotide at position 1 and the nucleotide at position 84 of SEQ ID No 4, or a biologically active fragment or variant thereof.

Another preferred 5'-regulatory polynucleotide of the invention includes the 5'-untranslated region (5'-UTR) located between the nucleotide at position 1 and the nucleotide at position 217 of SEQ ID No 5, or a biologically active fragment or variant thereof.

A first preferred 3'-regulatory polynucleotide of the invention includes a 3'-non coding region consisting in the nucleotide sequence starting from the nucleotide in position 988 and ending at the nucleotide in position 1414 of the nucleic acid of SEQ ID No 4, which is identical to the nucleotide sequence starting from the nucleotide in position 1121 and ending at the nucleotide in position 1547 of the nucleic acid of SEQ ID No 5. This first preferred 3'-regulatory polynucleotide carries a polyadenylation sit located between the nucleotide in position 1289 and the nucleotide in position 1294 of the nucleic acid of SEQ ID No 4 (and thus between the nucleotide in position 1422 and the nucleotide in position 1427 of the nucleic acid of SEQ ID No 5). Additionally, this first preferred 3'regulatory polynucleotide contains a potential polyadenylation site located between the nucleotide in position 1409 and the nucleotide in position 1414 of

the nucleic acid of SEQ ID No 4 (and thus between the nucleotide in position 1542 and the nucleotide in position 1547 of the nucleic acid of SEQ ID No 5).

A second preferred 3'-regulatory polynucleotide of the invention includes a 3'-non coding region consisting in the nucleotide sequence starting from the nucleotide in position 988 and ending at the nucleotide in position 1294 of the nucleic acid of SEQ ID No 4, which is identical to the nucleotide sequence starting from the nucleotide in position 1121 and ending at the nucleotide in position 1427 of the nucleic acid of SEQ ID No 5. This second preferred 3'-regulatory polynucleotide carries a polyadenylation site located between the nucleotide in position 1289 and the nucleotide in position 1294 of the nucleic acid of SEQ ID No 4 (and thus between the nucleotide in position 1422 and the nucleotide in position 1427 of the nucleic acid of SEQ ID No 5).

A further object of the invention consists of a purified or isolated nucleic acid comprising :

a) a nucleic acid comprising a regulatory polynucleotide of SEQ ID No 2 or a biologically active fragment or variant thereof;

b) a polynucleotide encoding a desired polypeptide or nucleic acid operably linked to the regulatory polynucleotide of SEQ ID No 2 or its biologically active fragment or variant thereof;

c) optionally, a nucleic acid comprising a regulatory polynucleotide of SEQ ID Nos 3 or a biologically active fragment or variant thereof.

In a specific embodiment of the nucleic acid defined above, said nucleic acid includes the 5'-untranslated region (5'-UTR) located between the nucleotide at position 1 and the nucleotide at position 84 of SEQ ID No 4, or a biologically active fragment or variant thereof.

In another specific embodiment of the nucleic acid defined above, said nucleic acid includes the 5'-untranslated region (5'-UTR) located between the nucleotide at position 1 and the nucleotide at position 217 of SEQ ID No 5, or a biologically active fragment or variant thereof.

In a third specific embodiment of the nucleic acid defined above, said nucl ic acid includes the 3'-untranslated region (3'-UTR) consisting in the nucleotide sequence starting from the nucleotide in position 988 and ending a the nucleotide in position 1414 of the nucleic acid of SEQ ID No 4.

In an additional preferred embodiment of the nucleic acid defined above, said nucleic acid includes the 3'-untranslated region (3'-UTR) consisting in the nucleotide

sequence starting from the nucleotide in position 988 and ending at the nucleotide in position 1294 of the nucleic acid of SEQ ID No 4.

The regulatory polynucleotide of SEQ ID No 2, or its biologically active fragments or variants, is advantageously located at the 5'-end of the polynucleotide encoding the desired polypeptide or polynucleotide.

The regulatory polynucleotide of SEQ ID No 3, or its biologically active fragments and variants, is advantageously placed at the 3'-end of the polynucleotide encoding the desired polypeptide or polynucleotide.

The desired polypeptide encoded by the above described nucleic acid may be of various nature or origin, encompassing proteins of prokaryotic or eukaryotic origin. Among the polypeptides expressed under the control of a *hGGPS* regulatory region, there may be cited bacterial, fungal or viral antigens. Also encompassed are eukaryotic proteins such as intracellular proteins, like "house keeping" proteins, membrane-bound proteins, like receptors, and secreted proteins like the numerous endogenous mediators such as cytokines.

The desired nucleic acids encoded by the above described polynucleotide, usually a RNA molecule, may be complementary to a desired coding polynucleotide, for example to the *hGGPS* coding sequence, and thus useful as an antisense polynucleotide.

Such a polynucleotide may be included in a recombinant expression vector in order to express the desired polypeptide or the desired nucleic acid in host cell or in a host organism. Suitable recombinant vectors that contain a polynucleotide such as described hereinbefore are disclosed elsewhere in the specification.

Coding regions

The *hGGPS* open reading frame is contained in the corresponding mRNAs of SEQ ID Nos 4 and 5.

More precisely, the effective *hGGPS* coding sequence (CDS) is comprised between the nucleotide at position 85 (first nucleotide of the ATG codon) and the nucleotide at position 987 (end nucleotide of the TAA codon) of SEQ ID No 4. A purified or isolated polynucleotide comprising the *hGGPS* coding region defined above is another object of the invention.

The above disclosed polynucleotide that contains the coding sequence of the *hGGPS* gene of the invention may be expressed in a desired host cell or a desired host organism, when this polynucleotide is placed under the control of suitable expression

signals. The expression signals may be either the expression signals contained in the regulatory regions in the *hGGPS* gene of the invention or in contrast be exogenous regulatory nucleic sequences. Such a polynucleotide, when placed under the suitabl expression signals, may also be inserted in a vector for its expression.

BIALLELIC MARKERS

The inventors have discovered nucleotide polymorphisms located within the genomic DNA containing the *hGGPS* gene, and among them "Single Nucleotide Polymorphisms" or SNPs that are also termed biallelic markers.

A) IDENTIFICATION OF BIALLELIC MARKERS

Biallelic markers consist of a single base polymorphism and are defined as genome-derived polynucleotides between 10 and 100, preferably between 20 30, or 40 and 60, more preferably about 45 nucleotides in length and most preferably 47 mer in length, which exhibit biallelic polymorphism at one single base position. Each biallelic marker therefore corresponds to two forms of a polynucleotide sequence included in a gene, which, when compared with one another, present a nucleotide modification at one position. Usually, the nucleotide modification involves the substitution of one nucleotide for another (for example A instead of T).

However, this nucleotide modification can also involve an insertion or a deletion of at least one nucleotide, preferably between 1 and 5 nucleotides. The nucleotide modification can also involve the presence of several adjacent single base polymorphisms. This type of nucleotide modification is usually called a "variable motif". Generally, a "variable motif" involves the presence of 2 to 10 adjacent single bas polymorphisms. In some instances, series of two or more single base polymorphisms can be interrupted by single bases which are not polymorphic. This is also globally considered to be a "variable motif".

Preferably, the lowest allele frequency of a biallelic polymorphism is 1%; sequence variants which show allele frequencies below 1% are called rare mutations. However, trait causing mutations may be present at a frequency less than 1%.

There are two preferred methods through which the biallelic markers of the present invention can be generated. In a first method, DNA samples from unrelated individuals are pooled together, following which the genomic DNA of interest is amplified and sequenced. The nucleotide sequences thus obtained are then analyzed to identify significant polymorphisms.

One of the major advantages of this method resides in the fact that the pooling of the DNA samples substantially reduces the number of DNA amplification reactions and sequencing reactions which must be carried out. Moreover, this method is sufficiently sensitive so that a biallelic marker obtained therewith usually shows a sufficient degree of informativeness for conducting association studies. The informative content of a biallelic marker contemplated by the present invention is preferably such that the frequency of its less frequent allele is not less than about 10 % (i.e. a heterozygosity rate of at least 0.18) (the heterozygosity rate for a biallelic marker is 2 P_a (1-P_a) , where P_a is the frequency of allele a). Preferably, the frequency of the less frequent allele of the biallelic markers contemplated within the invention is at least 20 % (i.e. a heterozygosity rate of at least 0.32). More preferably, the frequency of the less frequent allele of the biallelic markers contemplated within the invention is at least 30 % (i.e. its heterozygosity rate is higher than about 0.42).

In a second method for generating biallelic markers, the DNA samples are not pooled and are therefore amplified and sequenced individually. The resulting nucleotide sequences obtained are then also analyzed to identify significant polymorphisms.

It will readily be appreciated that when this second method is used, a substantially higher number of DNA amplification reactions and sequencing reactions must be carried out. Moreover, a biallelic marker obtained using this method may show a lower degree of informativeness for conducting association studies, e.g. if the frequency of its less frequent allele may be less than about 10%. Such a biallelic marker will however show a sufficient informative content to conduct association studies provided its less frequent allele is not less than about 0.01, i.e. its heterozygosity rate is higher than about 0.02. It will further be appreciated that including such less informative biallelic markers in association studies to identify potential genetic associations with a trait may allow in some cases the direct identification of causal mutations, which may, depending on their penetrance, be rare mutations. This method is usually preferred when biallelic markers need to be identified in order to perform association studies within candidate genes.

The following is a description of the various parameters of a preferred method used by the inventors to generate the markers of the present invention.

1. DNA extracti n

The genomic DNA samples from which the biallelic markers of the present invention are generated are preferably obtained from unrelated individuals corresponding to a heterogeneous population of known ethnic background.

The term "individual" as used herein refers to vertebrates, particularly members of the mammalian species and includes but is not limited to domestic animals, sports animals, laboratory animals, primates and humans. Preferably, the individual is a human.

The number of individuals from whom DNA samples are obtained can vary substantially, preferably from about 10 to about 1000, preferably from about 50 to about 200 individuals. It is usually preferred to collect DNA samples from at least about 100 individuals in order to have sufficient polymorphic diversity in a given population to identify as many markers as possible and to generate statistically significant results.

As for the source of the genomic DNA to be subjected to analysis, any test sample can be foreseen without any particular limitation. These test samples include biological samples which can be tested by the methods of the present invention described herein and include human and animal body fluids such as whole blood, serum, plasma, cerebrospinal fluid, urine, lymph fluids, and various external secretions of the respiratory, intestinal and genitourinary tracts, tears, saliva, milk, white blood cells, myelomas and the like; biological fluids such as cell culture supernatants; fixed tissue specimens including tumor and non-tumor tissue and lymph node tissues; bone marrow aspirates and fixed cell specimens. The preferred source of genomic DNA used in the context of the present invention is from peripheral venous blood of each donor.

The techniques of DNA extraction are well-known to the skilled technician. Such techniques are described notably by Linz et al. (1998) and by Mackey et al. (1998). Details of a preferred embodiment are provided in Example 2.

2. DNA amplification

DNA amplification techniques are well-known to those skilled in the art. Amplification techniques that can be used in the context of the present invention include, but are not limited to, the ligase chain reaction (LCR), the polymerase chain reaction (PCR, RT-PCR) and techniques such as the nucleic acid sequence based amplification (NASBA).

The primers according to the invention may be used in any of the following amplification procedures described below.

The PCR amplification reaction has been first described by Saiki et al. (1985). The Strand Displacement Amplification or SDA has been described by Walker et al., 1992. This amplification reaction is more completely disclosed in Spargo et al. (1996), which is herein incorporated by reference.

The Transcription-based Amplification System or TAS, as well as the Self-Sustained Sequence Replication or 3SR, the Nucleic Acid Sequence Based Amplification System or NASBA and also the Transcription Mediated Amplification or TMA are all amplification systems wherein the amplification reaction is conducted by an *in vitro* transcription reaction. TAS is described in details in Kwoh et al. (1989); 3SR is described in Guatelli et al. (1990); NASBA is described in Kievitis et al. (1991) and also by Bruisten et al. (1993) and Ovyn et al. (1996).

Other suitable techniques include the Ligase Chain Reaction or LCR (Landergren et al., 1988; Barany, 1991; European Patent Applications No EP-A-320 308 and EP-A-439 182), the Repair Chain Reaction or RCR (Segev et al., 1992), the Cycling Probe reaction or CPR (Duck et al., 1990) and the Qβ-replicase (Chu et al., 1986; Lizardi et al., 1988; Miele et al., 1983, Burg et al., 1996; Stone et al., 1996).

An amplification reaction technique encompassed by the present invention is described in Example 3.

The PCR technology is the preferred amplification technique used in the present invention. It has been described in several publications including US Patents 4,683,195, 4,683,202 and 4,965,188, the publication entitled "PCR Methods and Applications" (1991, Cold Spring Harbor Laboratory Press) and White et al. 1997. Each of these publications is incorporated by reference. A typical example of a PCR reaction suitable for the purposes of the present invention is provided in Example 3.

One of the aspects of the present invention is a method for the amplification of the *hGGPS* gene or a fragment or variant thereof in a test sample, preferably using the PCR technology. The method comprises the steps of contacting a test sample suspected of containing the target *hGGPS* encoding sequence or portion thereof with amplification reaction reagents comprising a pair of amplification primers, and eventually in some instances a detection probe that can hybridize with an internal region of amplicon sequences to confirm that the desired amplification reaction has taken place.

In this context, one of the groups of oligonucleotides according to the present invention is a first group of primers useful for the amplification of a genomic sequence encoding *hGGPS*. The primers pairs are characterized in that they have sufficient complementarity with any sequence of a strand of the *hGGPS* gene to be amplified, preferably with a sequence of introns adjacent to exons to amplify, with regions of the 3' and 5' ends of the *hGGPS* gene, with splice sites or with 5' UTRs or 3' UTRs to hybridize therewith.

These primers focus on exons and splice sites of the *hGGPS* gene since an identified biallelic marker as described below presents a higher probability to be an eventual causal mutation if it is located in these functional regions of the gene.

First primers and other oligonucleotides according to the invention are therefore synthesized to be "substantially" complementary to a strand of the *hGGPS* gene to be amplified. The primer sequence does not need to reflect the exact sequence of the DNA template. Minor mismatches can be accommodated by reducing the stringency of the hybridization conditions. Among the various methods available to design useful primers, the OSP computer software can be used by the skilled person (see Hillier & Green, 1991).

The first primers can be prepared by any suitable method, including, for example, cloning and restriction of appropriate sequences and direct chemical synthesis by a method such as the phosphodiester method of Narang et al. (1979), the phosphodiester method of Brown et al. (1979), the diethylphosphoramidite method of Beaucage et al. (1981) and the solid support method described in EP 0 707 592. The disclosures of all these documents are incorporated herein by reference.

The GC content in the first primers of the invention usually ranges between 10 and 75 %, preferably between 35 and 60 %, and more preferably between 40 and 55 %.

The length of the first primer can range from 10 to 100 nucleotides, preferably from 10 to 50, 10 to 30 or more preferably 10 to 25 nucleotides. Shorter primers tend to lack specificity for a target nucleic acid sequence and generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. Longer primers are expensive to produce and can sometimes self-hybridize to form hairpin structures. Preferred primers include those of SEQ ID Nos 9-10 described in Example 3. To these primers can be added, at either end thereof, a further polynucleotide useful for sequencing.

Other preferred primers according to the invention allow the amplification of various fragments of the purified or isolated nucleic acid of SEQ ID No 1. These primers are presented below as couples of forward and reverse primers that may be used together to amplify a desired nucleotide sequence.

a) : (1) Forward primer beginning at the nucleotide in position 7233 and ending a the nucleotide in position 7251 of the nucleotide sequence of SEQ ID No 1; (2) reverse primer which is complementary to the sequence beginning at the nucleotide in position 7565 and ending at the nucleotide in position 7582 of the nucleotide sequence of SEQ ID No 1.

b) : (1) Forward primer beginning at the nucleotide in position 13582 and ending a the nucleotide in position 13600 of the nucleotide sequence of SEQ ID No 1; (2) reverse primer which is complementary to the sequence beginning at the nucleotide in position 13982 and ending at the nucleotide in position 14001 of the nucleotide sequence of SEQ ID No 1.

c) : (1) Forward primer beginning at the nucleotide in position 14222 and ending a the nucleotide in position 14240 of the nucleotide sequence of SEQ ID No 1; (2) reverse primer which is complementary to the sequence beginning at the nucleotide in position 14626 and ending at the nucleotide in position 14645 of the nucleotide sequence of SEQ ID No 1.

d) : (1) Forward primer beginning at the nucleotide in position 14606 and ending a the nucleotide in position 14623 of the nucleotide sequence of SEQ ID No 1; (2) reverse primer which is complementary to the sequence beginning at the nucleotide in position 15007 and ending at the nucleotide in position 15026 of the nucleotide sequence of SEQ ID No 1.

e) : (1) Forward primer beginning at the nucleotide in position 14845 and ending a the nucleotide in position 14864 of the nucleotide sequence of SEQ ID No 1; (2) reverse primer which is complementary to the sequence beginning at the nucleotide in position 15246 and ending at the nucleotide in position 15265 of the nucleotid sequence of SEQ ID No 1.

The primers described above are individually useful as oligonucleotide probes in order to detect the corresponding *hGGPS* nucleotide sequence in a sample, and more preferably to detect the presence of a *hGGPS* DNA molecule in a sampl suspected to contain it.

3. Sequencing of amplified genomic DNA and identification f polymorphisms

The amplification products generated as described above with the primers of the invention are then sequenced using methods known and available to the skilled technician. Preferably, the amplified DNA is subjected to automated dideoxy terminator sequencing reactions using a dye-primer cycle sequencing protocol.

Following gel image analysis and DNA sequence extraction, sequence data are automatically processed with adequate software to assess sequence quality.

The sequence data obtained as described above are transferred to a proprietary database, where quality control and validation steps are performed. A proprietary base-caller ("Trace"), working using a Unix system automatically flags suspect peaks, taking into account the shape of the peaks, the inter-peak resolution, and the noise level. The proprietary base-caller also performs an automatic trimming. Any stretch of 25 or fewer bases having more than 4 suspect peaks is usually considered unreliable and is discarded.

After this first sequence quality analysis, polymorphism analysis software is used to detect the presence of biallelic sites among individual or pooled amplified fragment sequences. The polymorphism search is based on the presence of superimposed peaks in the electrophoresis pattern. These peaks, which present two distinct colors, correspond to two different nucleotides at the same position on the sequence. In order for peaks to be considered significant, peak height has to satisfy conditions of ratio between the peaks and conditions of ratio between a given peak and the surrounding peaks of the same color.

However, since the presence of two peaks can be an artifact due to background noise, two controls are utilized to exclude these artifacts :

- the two DNA strands are sequenced and a comparison between the peaks is carried out. The polymorphism has to be detected on both strands for validation.
- all the sequencing electrophoresis patterns of the same amplification product provided from distinct pools and/or individuals are compared. The homogeneity and the ratio of homozygous and heterozygous peak height are controlled through these distinct DNAs.

The detection limit for the frequency of biallelic polymorphisms detected by sequencing pools of 100 individuals is about 0.1 for the minor allele, as verified by sequencing pools of known allelic frequencies. However, more than 90 % of the

biallelic polymorphisms detected by the pooling method have a frequency for the minor allele higher than 0.25. Therefore, the biallelic markers selected by this method have a frequency of at least 0.1 for the minor allele and less than 0.9 for the major allele, preferably at least 0.2 for the minor allele and less than 0.8 for the major allele, more preferably at least 0.3 for the minor allele and less than 0.7 for the major allele, thus a heterozygosity rate higher than 0.18, preferably higher than 0.32, more preferably higher than 0.42.

In a particular embodiment of the invention, the test samples are a pool of 100 individuals and 50 individual samples. This is the methodology used in the preferred embodiment of the present invention, in which 1 biallelic marker has been identified in a genomic region containing the *hGGPS* gene.

The polymorphisms identified above can be further confirmed and their respective frequencies can be determined through various methods using the previously described primers and probes as described herein. These methods can also be useful for genotyping either new populations in association studies or individuals in the context of detection of alleles of biallelic markers which are known to be associated with a given trait. It will be appreciated that the methods described below can be equally performed on individual or pooled DNA samples.

B) GENOTYPING OF BIALLELIC MARKERS

Once a given polymorphic site has been found and characterized as a biallelic marker as described above, several methods can be used in order to determine the specific allele carried by an individual at the given polymorphic base.

The identification of biallelic markers described previously allows the design of appropriate oligonucleotides, which can be used as probes and primers, to amplify a *hGGPS* gene containing the polymorphic site of interest and for the detection of such polymorphisms.

1) Amplification

Most genotyping methods require the previous amplification of the DNA region carrying the polymorphic site of interest. Amplification can be performed using the same primers already detailed or alternative second primers.

The invention also concerns alternative second DNA primers, preferably in the form of primer pairs characterized in that they preferably comprise more than 8 nucleotides, more preferably between 8 and 25 nucleotides and in that they are

sufficiently complementary with a region of a *hGGPS* gene to hybridize therewith. In some embodiments, the primer pair is adapted for amplifying a sequence containing the polymorphic base of one of the sequences of SEQ ID Nos 7-8.

For amplification and sequencing, the pairs of primers are sufficiently complementary with a region of a *hGGPS* gene located at less than 500 bp, preferably at less than 100 bp, and more preferably at less than 50 bp of a polymorphic site corresponding to one of the markers of the present invention.

For allele specific amplification, at least one member of the pair of primers is sufficiently complementary with a region of a *hGGPS* gene comprising the polymorphic base in a biallelic marker of the present invention to hybridize therewith.

The GC content in the second primers of the invention usually ranges between 10 and 75 %, preferably between 35 and 60 %, and more preferably between 40 and 55 %.

The length of the primers of the present invention can range from 8 to 100 nucleotides, preferably from 8 to 50, 8 to 30 or more preferably 8 to 25 nucleotides. Shorter primers tend to lack specificity for a target nucleic acid sequence and generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. Longer primers are expensive to produce and can sometimes self-hybridize to form hairpin structures.

Methods for the synthesis of primers have been described previously and can be applied to the second primers of the invention.

One of the techniques that can be applied for the amplification of a polymorphic *hGGPS* gene or fragments thereof in a sample using the second primers of the invention can be selected from the techniques described above for the amplification of the *hGGPS* gene.

These second primers can be used, for example, for specific amplification experiments. In these experiments, primers which are complementary to a region of *hGGPS* DNA containing a biallelic marker are able to initiate the specific amplification of one allele of the biallelic marker.

2) Sequencing

The amplification products generated above with the primers of the invention can be sequenced using methods known and available to the skilled technician. Preferably, the amplified DNA is subjected to automated dideoxy terminator sequencing reactions using a

dye-primer cycle sequencing protocol. A sequence analysis can allow the identification of the base present at the polymorphic site.

3) Microsequencing

Polymorphism analyses on pools or selected individuals of a given population can be carried out by conducting microsequencing reactions on candidate regions contained in amplified fragments obtained by PCR performed on DNA or RNA samples taken from these individuals.

To do so, DNA samples are subjected to PCR amplification of the candidate regions under conditions similar to those described above. These genomic amplification products are then subjected to automated microsequencing reactions using ddNTPs (specific fluorescence for each ddNTP) and appropriate oligonucleotide microsequencing primers which can hybridize just upstream of the polymorphic base of interest. Once specifically extended at the 3' end by a DNA polymerase using a complementary fluorescent dideoxynucleotide analog (thermal cycling), the primer is precipitated to remove the unincorporated fluorescent ddNTPs. The reaction products in which fluorescent ddNTPs have been incorporated are then analyzed by electrophoresis on ABI 377 sequencing machines to determine the identity of the incorporated base, thereby identifying the polymorphic marker present in the sample.

An example of a typical microsequencing procedure that can be used in the context of the present invention is provided in example 5. It is to be understood that certain parameters of this procedure such as the electrophoresis method or the labeling of ddNTPs could be modified by the skilled person without substantially modifying its result.

Preferred microsequencing primers include the primer having the nucleotide sequence of SEQ ID No 11, as it is shown in Example 5.

As a further alternative to the process described above, several solid phas microsequencing reactions have been developed. The basic microsequencing protocol is the same as described previously, except that either the oligonucleotid microsequencing primers or the PCR-amplified products of the DNA fragment of interest are immobilized. For example, immobilization can be carried out via an interaction between biotinylated DNA and streptavidin-coated microtitration wells or avidin-coated polystyrene particles.

In such solid phase microsequencing reactions, incorporated ddNTPs can either be radiolabeled (see Syvänen, 1994, incorporated herein by reference) or linked

to fluorescein (see Livak & Hainer, 1994, incorporated herein by reference). The detection of radiolabeled ddNTPs can be achieved through scintillation-based techniques. The detection of fluorescein-linked ddNTPs can be based on the binding of antifluorescein antibody conjugated with alkaline phosphatase, followed by incubation with a chromogenic substrate (such as *p*-nitrophenyl phosphate).

Other possible of reporter-detection couples include :

- ddNTP linked to dinitrophenyl (DNP) and anti-DNP alkaline phosphatase conjugate (see Harju et al., 1993, incorporated herein by reference)
- biotinylated ddNTP and horseradish peroxidase-conjugated streptavidin with *o*-phenylenediamine as a substrate (see WO 92/15712, incorporated herein by reference).

A diagnosis kit based on fluorescein-linked ddNTP with antifluorescein antibody conjugated with alkaline phosphatase is commercialized under the name PRONTO by GamidaGen Ltd.

As yet another alternative microsequencing procedure, Nyren et al. (1993) presented a concept of solid-phase DNA sequencing that relies on the detection of DNA polymerase activity by an enzymatic luminometric inorganic pyrophosphate detection assay (ELIDA). The PCR-amplified products are biotinylated and immobilized on beads. The microsequencing primer is annealed and four aliquots of this mixture are separately incubated with DNA polymerase and one of the four different ddNTPs. After the reaction, the resulting fragments are washed and used as substrates in a primer extension reaction with all four dNTPs present. The progress of the DNA-directed polymerization reactions are monitored with the ELIDA. Incorporation of a ddNTP in the first reaction prevents the formation of pyrophosphate during the subsequent dNTP reaction. In contrast, no ddNTP incorporation in the first reaction gives extensive pyrophosphate release during the dNTP reaction and this leads to generation of light throughout the ELIDA reactions. From the ELIDA results, the first base after the primer is easily deduced.

Probes and primers

Nucleic acids of the invention that comprise at least 8 consecutive nucleotides of a nucleic acid selected from the group consisting of the nucleotide sequences SEQ ID Nos 2 and 3, the nucleotide sequences complementary thereto and the nucleotide

sequences hybridizing therewith under stringent hybridization conditions are all useful as polynucleotide probes or primers in order to detect the presence of a copy of the nucleic acid of SEQ ID No 1, as well as for detecting the presence of the corresponding mRNAs in a material sample.

Thus, the invention also relates to nucleic acid probes characterized in that they preferably comprise between 8 and 50 nucleotides that hybridize specifically, under the stringent hybridization conditions defined above, with a nucleic acid selected from the group consisting of the nucleotide sequences of SEQ ID Nos 2 and 3.

In a specific embodiment of the primers and probes according to the invention, these primers and probes comprise at least 8 consecutive nucleotides of a nucleic acid starting at the nucleotide in position 486 and ending at the nucleotide in position 7314 of the nucleotide sequence of SEQ ID No 2.

In another embodiment of the primers and probes according to the invention, these primers and probes have a length of at least 8 nucleotides and hybridize, under the stringent hybridization conditions defined above, with a nucleotide sequence found in a nucleic acid starting at the nucleotide in position 486 and ending at the nucleotide in position 7314 of the nucleotide sequence of SEQ ID No 2.

The GC content in the probes of the invention usually ranges between 10 and 75 %, preferably between 35 and 60 %, and more preferably between 40 and 55 %.

The length of these probes can range from 8, 10, 15, 20, or 30 to 100 nucleotides, preferably from 10 to 50, more preferably from 15 to 30 nucleotides. Shorter probes tend to lack specificity for a target nucleic acid sequence and generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. Longer probes are expensive to produce and can sometimes self-hybridize to form hairpin structures.

The primers and probes can be prepared by any suitable method, including, for example, cloning and restriction of appropriate sequences and direct chemical synthesis by a method such as the phosphodiester method of Narang et al. (1979), the phosphodiester method of Brown et al. (1979), the diethylphosphoramidite method of Beaucage et al. (1981) and the solid support method described in EP 0 707 592. The disclosures of all these documents are incorporated herein by reference.

The non-labeled probes of the invention may be directly used as probes. Nevertheless, the probes are preferably directly labeled such as with isotopes, reporter molecules or fluorescent labels or indirectly labeled such as with biotin to which a streptavidin complex may later bind. Probe labeling techniques are well-known to th

skilled technician. By assaying the presence of the probe, one can detect the presence or absence of the targeted DNA sequence in a given sample. The same labels can be used with primers.

The probes are generally labeled with a radioactive element (^{32}P, ^{35}S, ^{3}H, ^{125}I) or by a non-isotopic molecule (for example, biotin, acetylaminofluorene, digoxigenin, 5-bromodesoxyuridin, fluorescein).

Examples of non-radioactive labeling of nucleic acid fragments are described in the French patent N° FR-7810975 or by Urdea et al (1988) or Sanchez-Pescador et al (1988)

Advantageously, the probes according to the present invention may have structural characteristics such that they allow the signal amplification, such structural characteristics being, for example, branched DNA probes as those described by Urdea et al. in 1991 or in the European patent N° EP-0225,807 (Chiron).

The probes of the present invention are useful for a number of purposes. They can be notably used in Southern hybridization to genomic DNA. The probes can also be used to detect PCR amplification products. They may also be used to detect mismatches in the *hGGPS* gene or mRNA using other techniques. Generally, the probes are complementary to the *hGGPS* regulatory sequences.

Any of the primers and probes of the present invention can be conveniently immobilized on a solid support. Solid supports are known to those skilled in the art and include the walls of wells of a reaction tray, test tubes, polystyrene beads, magnetic beads, nitrocellulose strips, membranes, microparticles such as latex particles, sheep (or other animal) red blood cells, duracytes and others. The "solid phase" is not critical and can be selected by one skilled in the art. Thus, latex particles, microparticles, magnetic or non-magnetic beads, membranes, plastic tubes, walls of microtiter wells, glass or silicon chips, sheep (or other suitable animal's) red blood cells and duracytes are all suitable examples.

Suitable methods for immobilizing nucleic acids on solid phases include ionic, hydrophobic, covalent interactions and the like. A "solid phase", as used herein, refers to any material which is insoluble, or can be made insoluble by a subsequent reaction. The solid phase can be chosen for its intrinsic ability to attract and immobilize the capture reagent.

Alternatively, the solid phase can retain an additional receptor which has the ability to attract and immobilize the capture reagent. The additional receptor can

include a charged substance that is oppositely charged with respect to the capture reagent itself or to a charged substance conjugated to the capture reagent.

As yet another alternative, the receptor molecule can be any specific binding member which is immobilized upon (attached to) the solid phase and which has the ability to immobilize the capture reagent through a specific binding reaction. The receptor molecule enables the indirect binding of the capture reagent to a solid phase material before the performance of the assay or during the performance of the assay. The solid phase thus can be a plastic, derivatized plastic, magnetic or non-magnetic metal, glass or silicon surface of a test tube, microtiter well, sheet, bead, microparticle, chip, sheep (or other suitable animal's) red blood cells, duracytes and other configurations known to those of ordinary skill in the art.

Consequently, the invention also deals with a method for detecting the presence of a nucleic acid comprising at least a part of a nucleotide sequence selected from the group consisting of SEQ ID Nos 2, 3 and 7-11 in a sample, said method comprising the following steps of :

a) bringing into contact a nucleic acid probe or a plurality of nucleic acid probes, which can hybridize to a nucleotide sequence included in one of the nucleic acids of SEQ ID Nos 2, 3 and 7-11, and the sample to be assayed.

b) detecting the hybrid complex formed between the probe and a nucleic acid in the sample.

Preferably, the nucleic acid probe is selected from the group of polynucleotides consisting of the nucleotide sequences SEQ ID Nos 7-11. In a first preferred embodiment of this detection method, said nucleic acid probe or the plurality of nucleic acid probes are labeled with a detectable molecule. In a second preferred embodiment of said method, said nucleic acid probe or the plurality of nucleic acid probes has been immobilized on a substrate.

The invention further concerns a kit for detecting the presence of a nucleic acid comprising at least a part of a nucleotide sequence selected from the group consisting of SEQ ID Nos 2, 3 and 7-11 in a sample, said kit comprising :

a) a nucleic acid probe or a plurality of nucleic acid probes which can hybridize to a nucleotide sequence included in one of the nucleic acids of SEQ ID Nos 2, 3 and 7-11;

b) optionally, the reagents necessary for performing the hybridization reaction.

The nucleic acid probe or the plurality of nucleic acid probes that are included in the detection kit described above may be selected from the group consisting of SEQ ID Nos 9-11. In a first preferred embodiment of the detection kit, the nucleic acid probe or the plurality of nucleic acid probes are labeled with a detectable molecule. In a second preferred embodiment of the detection kit, the nucleic acid probe or the plurality of nucleic acid probes has been immobilized on a substrate.

Oligonucleotide arrays

An oligonucleotide probe matrix may advantageously be used to detect mutations occurring in the *hGGPS* gene and preferably in its regulatory regions. For this particular purpose, probes are specifically designed to have a nucleotide sequence allowing their hybridization to the genes that carry known mutations (either by deletion, insertion of substitution of one or several nucleotides). By known mutations is meant mutations on the *hGGPS* gene that have been identified according, for example to the technique used by Huang et al. (1996) or Samson et al. (1996).

Another technique that is used to detect mutations in the *hGGPS* gene is the use of a high-density DNA array. Each oligonucleotide probe constituting a unit element of the high density DNA array is designed to match a specific subsequence of the *hGGPS* genomic DNA or cDNA. Thus, an array consisting of oligonucleotides complementary to subsequences of the target gene sequence is used to determine the identity of the target sequence with the wild gene sequence, measure its amount, and detect differences between the target sequence and the reference wild gene sequence of the *hGGPS* gene. In one such design, termed 4L tiled array, is implemented a set of four probes (A, C, G, T), preferably 15-nucleotide oligomers. In each set of four probes, the perfect complement will hybridize more strongly than mismatched probes. Consequently, a nucleic acid target of length L is scanned for mutations with a tiled array containing 4L probes, the whole probe set containing all the possible mutations in the known wild reference sequence. The hybridization signals of the 15-mer probe set tiled array are perturbed by a single base change in the target sequence. As a consequence, there is a characteristic loss of signal or a « footprint » for the probes flanking a mutation position. This technique was described by Chee et al. in 1996, which is herein incorporated by reference.

Vectors for the expression of a regulatory or a coding polynucleotide according to the invention.

Any of the regulatory polynucleotides or the coding polynucleotides of the invention may be inserted into recombinant vectors for expression in a recombinant host cell or a recombinant host organism.

Thus, the present invention also encompasses a family of recombinant vectors that contains either a regulatory polynucleotide selected from the group consisting of the regulatory polynucleotides derived from the *hGGPS* gene, or a polynucleotide comprising the *hGGPS* coding sequence, or both.

More particularly, the present invention relates to expression vectors which include nucleic acids encoding the hGGPS protein of the amino acid sequence of SEQ ID No 6 described therein under the control of either one regulatory sequence selected among the *hGGPS* regulatory polynucleotides, or alternatively under the control of an exogenous regulatory sequence.

A recombinant expression vector comprising a nucleic acid selected from the group consisting of SEQ ID Nos 2 and 3, or biologically active fragments or variants thereof, is also part of the present invention.

The invention also encompasses a recombinant expression vector comprising :

a) a nucleic acid comprising a regulatory polynucleotide of the nucleotide sequence SEQ ID Nos 2, or a biologically active fragment or variant thereof;

b) a polynucleotide encoding a polypeptide or a polynucleotide of interest.

c) optionally, a nucleic acid comprising a regulatory polynucleotide of SEQ ID No 3 or a biologically active fragment or variant thereof.

The invention also pertains to a recombinant expression vector useful for th expression of the *hGGPS* coding sequence, wherein said vector comprises a nucleic acid selected from the group of SEQ ID Nos 1, 4 and 5 or a nucleic acid having at least 95% nucleotide identity with a polynucleotide selected from the group consisting of the nucleotide sequences of SEQ ID Nos 1, 4 and 5.

Another recombinant expression vector of the invention consists in a recombinant vector comprising a nucleic acid comprising the nucleotide sequence beginning at the nucleotide in position 85 and ending in position 987 of the polynucleotide of SEQ ID No 4.

Some of the elements which can be found in the vectors of the present invention are described in further detail in the following sections.

a) Vectors

A recombinant vector according to the invention comprises, but is not limited to, a YAC (Yeast Artificial Chromosome), a BAC (Bacterial Artificial Chromosome), a phage, a phagemid, a cosmid, a plasmid or even a linear DNA molecule which may consist of a chromosomal, non-chromosomal and synthetic DNA. Such a recombinant vector can comprise a transcriptional unit comprising an assembly of :

(1) a genetic element or elements having a regulatory role in gene expression, for example promoters or enhancers. Enhancers are cis-acting elements of DNA, usually from about 10 to 300 bp in length that act on the promoter to increase the transcription.

(2) a structural or coding sequence which is transcribed into mRNA and eventually translated into a polypeptide, and

(3) appropriate transcription initiation and termination sequences. Structural units intended for use in yeast or eukaryotic expression systems preferably include a leader sequence enabling extracellular secretion of translated protein by a host cell. Alternatively, where recombinant protein is expressed without a leader or transport sequence, it may include an N-terminal residue. This residue may or may not be subsequently cleaved from the expressed recombinant protein to provide a final product.

Generally, recombinant expression vectors will include origins of replication, selectable markers permitting transformation of the host cell, and a promoter derived from a highly expressed gene to direct transcription of a downstream structural sequence. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium.

The selectable marker genes for selection of transformed host cells are preferably dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, TRP1 for *S. cerevisiae* or tetracycline, rifampicin or ampicillin resistance in *E. coli*, or levan saccharase for mycobacteria.

As a representative but non-limiting example, useful expression vectors for bacterial use can comprise a selectable marker and bacterial origin of replication derived from commercially available plasmids comprising genetic elements of pBR322

(ATCC 37017). Such commercial vectors include, for example, pKK223-3 (Pharmacia, Uppsala, Sweden), and GEM1 (Promega Biotec, Madison, WI, USA).

Large numbers of suitable vectors and promoters are known to those of skill in the art, and commercially available, such as bacterial vectors : pQE70, pQE60, pQE-9 (Qiagen), pbs, pD10, phagescript, psiX174, pbluescript SK, pbsks, pNH8A, pNH16A, pNH18A, pNH46A (Stratagene); ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia); or eukaryotic vectors : pWLNEO, pSV2CAT, pOG44, pXT1, pSG (Stratagene); pSVK3, pBPV, pMSG, pSVL (Pharmacia); baculovirus transfer vector pVL1392/1393 (Pharmingen); pQE-30 (QIAexpress).

A suitable vector for the expression of the hGGPS polypeptide of SEQ ID No 12 is a baculovirus vector that can be propagated in insect cells and in insect cell lines. A specific suitable host vector system is the pVL1392/1393 baculovirus transfer vector (Pharmingen) that is used to transfect the SF9 cell line (ATCC N°CRL 1711) which is derived from *Spodoptera frugiperda*.

Other suitable vectors for the expression of the hGGPS polypeptide of SEQ ID No12 in a baculovirus expression system include those described by Chai et al. (1993), Vlasak et al. (1983) and Lenhard et al. (1996).

Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA sequences derived from the SV40 viral genome, for example SV40 origin, early promoter, enhancer, splice and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

b) Promoters

The suitable promoter regions used in the expression vectors according to the present invention are chosen taking into account the cell host in which the heterologous gene has to be expressed.

A suitable promoter may be heterologous with respect to the nucleic acid for which it controls the expression or alternatively can be endogenous to the native polynucleotide containing the coding sequence to be expressed. Additionally, the promoter is generally heterologous with respect to the recombinant vector sequences within which the construct promoter/coding sequence has been inserted.

Preferred bacterial promoters are the LacI, LacZ, the T3 or T7 bacteriophage RNA polymerase promoters, the polyhedrin promoter, or the p10 protein promoter from baculovirus (Kit Novagen) (Smith et al., 1983; O'Reilly et al., 1992), the lambda P_R promoter or also the trc promoter.

Promoter regions can be selected from any desired gene using, for example, CAT (chloramphenicol transferase) vectors and more preferably pKK232-8 and pCM7 vectors. Particularly preferred bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda PR, PL and trp. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-L. Selection of a convenient vector and promoter is well within the level of ordinary skill in the art.

The choice of a promoter is well within the ability of a person skilled in the field of genetic egineering. For example, one may refer to the book of Sambrook et al. (1989) or also to the procedures described by Fuller et al. (1996).

The vector containing the appropriate DNA sequence as described above, more preferably a *hGGPS* gene regulatory polynucleotide, a polynucleotide encoding the hGGPS polypeptide of SEQ ID No 6 or both of them, can be utilized to transform an appropriate host to allow the expression of the desired polypeptide or polynucleotide.

c) Other types of vectors

The *in vivo* expression of a hGGPS polypeptide of SEQ ID No 6 may be useful in order to correct a genetic defect related to the expression of the native gene in a host organism or to the production of a biologically inactive hGGPS protein.

Consequently, the present invention also deals with recombinant expression vectors mainly designed for the *in vivo* production of the hGGPS polypeptide of SEQ ID No 6 by the introduction of the appropriate genetic material in the organism of the patient to be treated. This genetic material may be introduced *in vitro* in a cell that has been previously extracted from the organism, the modified cell being subsequently reintroduced in the said organism, directly *in vivo* into the appropriate tissue, and preferably in the olfactory epithelium.

By « vector » according to this specific embodiment of the invention is intended either a circular or a linear DNA molecule.

One specific embodiment for a method for delivering a protein or peptide to the interior of a cell of a vertebrate *in vivo* comprises the step of introducing a preparation

comprising a physiologically acceptable carrier and a naked polynucleotide operatively coding for the polypeptide of interest into the interstitial space of a tissue comprising the cell, whereby the naked polynucleotide is taken up into the interior of the cell and has a physiological effect.

In a specific embodiment, the invention provides a composition for the *in vivo* production of the hGGPS protein or polypeptide described herein. It comprises a naked polynucleotide operatively coding for this polypeptide, in solution in a physiologically acceptable carrier, and suitable for introduction into a tissue to cause cells of the tissue to express the said protein or polypeptide.

Compositions comprising a polynucleotide are described in the PCT application N° WO 90/11092 (Vical Inc.) and also in the PCT application N° WO 95/11307 (Institut Pasteur, INSERM, Université d'Ottawa) as well as in the articles of Tacson et al. (1996) and of Huygen et al. (1996).

The amount of the vector to be injected to the desired host organism vary according to the site of injection. As an indicative dose, it will be injected between 0,1 and 100 µg of the vector in an animal body, preferably a mammal body, for example a mouse body.

In another embodiment of the vector according to the invention, it may be introduced in vitro in a host cell, preferably in a host cell previously harvested from the animal to be treated and more preferably a somatic cell such as a muscle cell. In a subsequent step, the cell that has been transformed with the vector coding for the desired hGGPS polypeptide or the desired C-terminal fragment thereof is reintroduced into the animal body in order to deliver the recombinant protein within the body either locally or systemically.

In one specific embodiment, the vector is derived from an adenovirus. Preferred adenovirus vectors according to the invention are those described by Feldman and Steg (1996) or Ohno et al. (1994). Another preferred recombinant adenovirus according to this specific embodiment of the present invention is the human adenovirus type 2 or 5 (Ad 2 or Ad 5) or an adenovirus of animal origin (French patent application N° FR-93.05954).

Retrovirus vectors and adeno-associated virus vectors are generally understood to be the recombinant gene delivery system of choice for the transfer of exogenous polynucleotides *in vivo* , particularly to mammals, including humans. These vectors provide efficient delivery of genes into cells, and the transferred nucleic acids are stably integrated into the chromosomal DNA of the host

Particularly preferred retroviruses for the preparation or construction of retroviral *in vitro* or *in vitro* gene delivery vehicles of the present invention include retroviruses selected from the group consisting of Mink-Cell Focus Inducing Virus, Murine Sarcoma Virus, Reticuloendotheliosis virus and Rous Sarcoma virus. Particularly preferred Murine Leukemia Viruses include the 4070A and the 1504A viruses, Abelson (ATCC No VR-999), Friend (ATCC No VR-245), Gross (ATCC No VR-590), Rauscher (ATCC No VR-998) and Moloney Murine Leukemia Virus (ATCC No VR-190; PCT Application No WO 94/24298). Particularly preferred Rous Sarcoma Viruses include Bryan high titer (ATCC Nos VR-334, VR-657, VR-726, VR-659 and VR-728). Other preferred retroviral vectors are those described in Roth et al. (Roth J.A. et al., 1996), the PCT Application No WO 93/25234, the PCT Application No WO 94/06920, Roux et al., 1989, Julan et al., 1992 and Neda et al., 1991.

Yet another viral vector system that is contemplated by the invention consists in the adeno-associated virus (AAV). The adeno-associated virus is a naturally occurring defective virus that requires another virus, such as an adenovirus or a herpes virus, as a helper virus for efficient replication and a productive life cycle (Muzyczka et al., 1992). It is also one of the few viruses that may integrate its DNA into non-dividing cells, and exhibits a high frequency of stable integration (Flotte et al., 1992; Samulski et al., 1989; McLaughlin et al., 1989). One advantageous feature of AAV derives from its reduced efficacy for transducing primary cells relative to transformed cells.

Other compositions containing a vector of the invention advantageously comprise an oligonucleotide fragment of a nucleic sequence selected from the group consisting of SEQ ID Nos 2 or 3 as an antisense tool that inhibits the expression of the corresponding *hGGPS* gene. Preferred methods using antisense polynucleotide according to the present invention are the procedures described by Sczakiel et al. (Sczakiel G. et al., 1995, Trends Microbiol., 1995, Vol. 3(6):213-217) or also in the PCT Application No WO 95/24223.

Preferably, the antisense tools are chosen among the polynucleotides (15-200 bp long) that are complementary to the 5'end of the *hGGPS* mRNAs. In another embodiment, a combination of different antisense polynucleotides complementary to different parts of the desired targeted gene are used.

Preferred antisense polynucleotides according to the present invention are complementary to a sequence of the mRNAs of *hGGPS* that contains the translation initiation codon ATG.

Host cells

Another object of the invention consists in cell host that have been transformed or transfected with one of the polynucleotides described therein, and more precisely a polynucleotide either comprising a *hGGPS* regulatory polynucleotide or the coding sequence of the hGGPS polypeptide having the amino acid sequence of SEQ ID No 6. Are included cell hosts that are transformed (prokaryotic cells) or that are transfected (eukaryotic cells) with a recombinant vector such as those described above.

A cell host according to the present invention is characterized in that its genome or genetic background (including chromosome, plasmids) is modified by the heterologous nucleic acid coding for the hGGPS polypeptide of SEQ ID No 6.

Preferred cell hosts used as recipients for the expression vectors of th invention are the following :

a) Prokaryotic host cells : *Escherichia coli* strains (I.E. DH5-α strain) or *Bacillus subtilis*.

b) Eukaryotic host cells : HeLa cells (ATCC N°CCL2; N°CCL2.1; N°CCL2.2), Cv 1 cells (ATCC N°CCL70), COS cells (ATCC N°CRL1650; N°CRL1651), Sf-9 cells (ATCC N°CRL1711).

The constructs in the host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence.

Following transformation of a suitable host and growth of the host to an appropriate cell density, the selected promoter is induced by appropriate means, such as temperature shift or chemical induction, and cells are cultivated for an additional period.

Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification.

Microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents. Such methods are well known by the skill artisan.

The hGGPS polypeptide of SEQ ID No 6.

It is now routine to produce proteins in high amounts with genetic engineering techniques through the use, as expression vectors, of plasmids, phages or phagemids. One of the polynucleotides that code for the polypeptides of the present invention is

inserted in an appropriate expression vector, such as those described above, in order to produce *in vitro* the polypeptide of interest.

Thus, the invention also pertains to the hGGPS polypeptide having the amino acid sequence of SEQ ID No 6 for example as produced by the genetic engineering techniques.

The invention also concerns the hGGPS polypeptide encoded by one of the polynucleotides selected from the group consisting of the nucleotide sequences of SEQ ID Nos 4 and 5.

The polypeptides according to the invention may also be prepared by the conventional methods of chemical synthesis, either in a homogenous solution or in solid phase. As an illustrative embodiment of such chemical polypeptide synthesis techniques, it may be cited the homogenous solution technique described by Houbenweyl in 1974. For solid phase synthesis the technique described by Merrifield (1965) may be used in particular.

Antibodies

The polypeptide of SEQ ID No 6 can be used for the preparation of polyclonal or monoclonal antibodies.

The hGGPS polypeptide expressed from a DNA sequence comprising at least one of the nucleic acid sequences of SEQ ID Nos 1, 4 and 5 may also be used to generate antibodies capable of specifically binding to the hGGPS polypeptide of SEQ ID No 6.

In a preferred embodiment of polyclonal or monoclonal antibodies of the invention consists in antibodies raised against a C-terminal portion of the hGGPS polypeptide of the amino acid sequence of SEQ ID No 6, more preferably antibodies raise against a peptide fragment of the hGGPS polypeptide having the amino acid sequence starting from the amino acid at position 200 and ending at the amino acid in position 300 of the hGGPS polypeptide of SEQ ID No 6, or peptide fragments th reof.

The antibodies may be prepared from hybridomas according to the technique described by Kohler and Milstein in 1975. The polyclonal antibodies may be prepared by immunization of a mammal, especially a mouse or a rabbit, with a polypeptide according to the invention that is combined with an adjuvant of immunity, and then by purifying of the specific antibodies contained in the serum of the immunized animal on a affinity chromatography column on which has previously been immobilized the polypeptide that has been used as the antigen.

The present invention also includes, chimeric single chain Fv antibody fragments (Martineau et al., 1998), antibody fragments obtained through phage display libraries (Ridder et al., 1995; Vaughan et al., 1995) and humanized antibodies (Reinmann et al., 1997; Leger et al., 1997).

Methods and kits for screening candidate substances or molecules modulating the expression of the* hGGPS *gene.

The present invention also concerns a method for screening substances or molecules that are able to increase, or in contrast to decrease or even to suppress, the expression of the *hGGPS* gene. Such a method may allow one skilled in the art to select substances exerting a regulating effect on the expression level of the *hGGPS* gene and thus enabling a correction in the *hGGPS* expression levels in individuals in which the *hGGPS* expression is defective (i.e. lower or in contrast higher than the normal expression levels).

The alteration of the *hGGPS* expression in response to a modifier can be determined by administering or combining the candidate expression modifier with an expression system such as animals, cells, and *in vitro* transcription assays.

The term "expression modifier" is intended to encompass but is not limited to chemical agents that modulate the *hGGPS* gene expression.

The effect of the modifier on *hGGPS* transcription and /or steady state mRNA levels can be also determined. As it is the case for basic expression levels, tissu specific interactions are of interest. A panel of different modifiers may be screened in order to determine the effect under a number of different conditions.

The screening of modifiers can also be carried out with a construct which comprises the regulatory region of the *hGGPS* gene or a portion thereof operably linked to a reporter gene such as luciferase, β galactosidase, green fluorescent protein (GFP) and chloramphenicol acetyl transferase (CAT).

Hybridization with long probes

Expression levels and patterns of *hGGPS* may be analyzed by solution hybridization with long probes as described in International Patent Application No. WO 97/05277, the entire contents of which are hereby incorporated by reference. Briefly, the *hGGPS* genomic DNA described above, more particularly a sequence selected from the group consisting of SEQ ID Nos 1-5 or fragments thereof, is inserted at a cloning site immediately downstream of a bacteriophage (T3, T7 or SP6) RNA polymerase promoter to produce antisense RNA. Preferably, the *hGGPS* insert

comprises at least 100 or more consecutive nucleotides of the genomic DNA sequence and most preferably of the genomic sequence contained in the nucleotide sequences of SEQ ID Nos 2 and 3. The plasmid is linearized and transcribed in the presence of ribonucleotides comprising modified ribonucleotides (i.e. biotin-UTP and DIG-UTP). An excess of this doubly labeled RNA is hybridized in solution with mRNA isolated from cells or tissues of interest. The hybridizations are performed under standard stringent conditions (40-50°C for 16 hours in an 80% formamide, 0.4 M NaCl buffer, pH 7-8). The unhybridized probe is removed by digestion with ribonucleases specific for single-stranded RNA (i.e. RNases CL3, T1, Phy M, U2 or A). The presence of the biotin-UTP modification enables the capture of the hybrid on a microtitration plate coated with streptavidin. The presence of the DIG modification enables the hybrid to be detected and quantified by ELISA using an anti-DIG antibody coupled to alkaline phosphatase.

Assays

Quantitative analysis of *hGGPS* gene expression may also be performed using arrays. As used herein, the term array means a one dimensional, two dimensional, or multidimensional arrangement of a plurality of nucleic acids of sufficient length to permit specific detection of expression of mRNAs capable of hybridizing thereto. For example, the arrays may contain a plurality of nucleic acids derived from genes whose expression levels are to be assessed. The arrays may include the *hGGPS* genomic DNA, or sequences complementary thereto or fragments thereof. Preferably, the array includes nucleotide sequences that are comprised in the non coding 5'-UTR or the non coding 3'-UTR of the *hGGPS* cDNAS of SEQ ID Nos 4 or 5, and most preferably nucleotide sequences located at the 3'end of the nucleic acid of SEQ ID Nos 2 and 3 or alternatively nucleotide sequences located at the 5'-end of the nucleic acid of SEQ ID No 4. Preferably, the fragments are at least 15 nucleotides in length. In other embodiments, the fragments are at least 25 nucleotides in length. In some embodiments, the fragments are at least 50 nucleotides in length. More preferably, the fragments are at least 100 nucleotides in length. In another preferred embodiment, the fragments are more than 100 nucleotides in length. In some embodiments the fragments may be more than 500 nucleotides in length.

For example, quantitative analysis of *hGGPS* gene expression may be performed with a cDNA microarray as described by Schena et al. (1995 and 1996). Full length *hGGPS* cDNAs or fragments thereof are amplified by PCR and arrayed from a 96-well microtiter plate onto silylated microscope slides using high-speed robotics.

Printed arrays are incubated in a humid chamber to allow rehydration of the array elements and rinsed, once in 0.2% SDS for 1 min, twice in water for 1 min and once for 5 min in sodium borohydride solution. The arrays are submerged in water for 2 min at 95°C, transferred into 0.2% SDS for 1 min, rinsed twice with water, air dried and stored in the dark at 25°C.

Cell or tissue mRNA is isolated or commercially obtained and probes are prepared by a single round of reverse transcription. Probes are hybridized to 1 cm^2 microarrays under a 14 x 14 mm glass coverslip for 6-12 hours at 60°C. Arrays are washed for 5 min at 25°C in low stringency wash buffer (1 x SSC/0.2% SDS), then for 10 min at room temperature in high stringency wash buffer (0.1 x SSC/0.2% SDS). Arrays are scanned in 0.1 x SSC using a fluorescence laser scanning device fitted with a custom filter set. Accurate differential expression measurements are obtained by taking the average of the ratios of two independent hybridizations.

Quantitative analysis of the *hGGPS* gene expression may also be performed with full length *hGGPS* cDNAs or fragments thereof in complementary DNA arrays as described by Pietu et al. (1996). The full length *hGGPS* cDNA or fragments thereof is PCR amplified and spotted on membranes. Then, mRNAs originating from various tissues or cells are labeled with radioactive nucleotides. After hybridization and washing in controlled conditions, the hybridized mRNAs are detected by phospho-imaging or autoradiography. Duplicate experiments are performed and a quantitative analysis of differentially expressed mRNAs is then performed.

Alternatively, expression analysis using the *hGGPS* genomic DNA sequences, or fragments thereof can be done through high density nucleotide arrays as described by Lockhart et al. (1996) and Sosnowsky et al. (1997). Oligonucleotides of 15-50 nucleotides from the sequences of the *hGGPS* genomic DNA sequences or the sequences complementary thereto, are synthesized directly on the chip (Lockhart t al., supra) or synthesized and then addressed to the chip (Sosnowski et al., supra). Preferably, the oligonucleotides are about 20 nucleotides in length.

hGGPS cDNA probes labeled with an appropriate compound, such as biotin, digoxigenin or fluorescent dye, are synthesized from the appropriate mRNA population and then randomly fragmented to an average size of 50 to 100 nucleotides. These probes are then hybridized to the chip. After washing as described in Lockhart et al., supra and application of different electric fields (Sosnowsky et al., 1997), the dyes or labeling compounds are detected and quantified. Duplicate hybridizations are

performed. Comparative analysis of the intensity of the signal originating from cDNA probes on the same target oligonucleotide in different cDNA samples indicates a differential expression of the *hGGPS* mRNAs.

EXAMPLES

Example 1 : Analysis of the mRNAs encoding the hGGPS polypeptide of SEQ ID No 6 synthesized by the cells.

Human *GGPS* cDNA was obtained as follows : 4µl of ethanol suspension containing 1 mg of human prostate total RNA (Clontech laboratories, Inc., Palo Alto, USA; Catalogue N. 64038-1) was centrifuged, and the resulting pellet was air dried for 30 minutes at room temperature.

First strand cDNA synthesis was performed using the AdvantageTM RT-for-PCR kit (Clontech laboratories Inc., catalogue N. K1402-1). 1 µl of 20 mM solution of a specific oligo dT primer was added to 12.5 µl of RNA solution in water, heated at 74°C for 2.5 min and rapidly quenched in an ice bath. 10 µl of 5 x RT buffer (50 mM Tris-HCl, pH 8.3, 75 mM KCl, 3 mM $MgCl_2$), 2.5 µl of dNTP mix (10 mM each), 1.25 µl of human recombinant placental RNA inhibitor were mixed with 1 ml of MMLV reverse transcriptase (200 units). 6.5 µl of this solution were added to RNA-primer mix and incubated at 42°C for one hour. 80 µl of water were added and the solution was incubated at 94°C for 5 minutes.

5µl of the resulting solution were used in a Long Range PCR reaction with hot start, in 50 µl final volume, using 2 units of rtTHXL, 20 pmol/µl of each of 5'-TGGAGAAGACTCAAGAAACAGTCCAAA-3' (from the nucleotide in position 86 to the nucleotide in position 112 of SEQ ID No 4) and 5'-CCTGGAAGCAAGTCTTTTTTATTGACG-3' (from the nucleotide in position 1285 to the nucleotide in position 1311 of SEQ ID No 4) primers with 35 cycles of elongation for 6 minutes at 67°C in thermocycler.

The amplification products corresponding to both cDNA strands are partially sequenced in order to ensure the specificity of the amplification reaction.

Results of Northern blot analysis of prostate mRNAs support the existence of a *hGGPS* cDNA which corresponds to the nucleotide sequence of SEQ ID No 4.

Example 2 : Detection of *hGGPS* biallelic markers: DNA extraction

Donors were unrelated and healthy. They presented a sufficient diversity for being representative of a French heterogeneous population. The DNA from 100 individuals was extracted and tested for the detection of the biallelic markers.

30 ml of peripheral venous blood were taken from each donor in the presence of EDTA. Cells (pellet) were collected after centrifugation for 10 minutes at 2000 rpm. Red cells were lysed by a lysis solution (50 ml final volume : 10 mM Tris pH7.6; 5 mM $MgCl_2$; 10 mM NaCl). The solution was centrifuged (10 minutes, 2000 rpm) as many times as necessary to eliminate the residual red cells present in the supernatant, after resuspension of the pellet in the lysis solution.

The pellet of white cells was lysed overnight at 42°C with 3.7 ml of lysis solution composed of:

- 3 ml TE 10-2 (Tris-HCl 10 mM, EDTA 2 mM) / NaCl 0.4 M
- 200 µl SDS 10%
- 500 µl K-proteinase (2 mg K-proteinase in TE 10-2 / NaCl 0.4 M).

For the extraction of proteins, 1 ml saturated NaCl (6M) (1/3.5 v/v) was added. After vigorous agitation, the solution was centrifuged for 20 minutes at 10000 rpm.

For the precipitation of DNA, 2 to 3 volumes of 100% ethanol were added to the previous supernatant, and the solution was centrifuged for 30 minutes at 2000 rpm. The DNA solution was rinsed three times with 70% ethanol to eliminate salts, and centrifuged for 20 minutes at 2000 rpm. The pellet was dried at 37°C, and resuspended in 1 ml TE 10-1 or 1 ml water. The DNA concentration was evaluated by measuring the OD at 260 nm (1 unit OD = 50 µg/ml DNA).

To determine the presence of proteins in the DNA solution, the OD 260 / OD 280 ratio was determined. Only DNA preparations having a OD 260 / OD 280 ratio between 1.8 and 2 were used in the subsequent examples described below.

The pool was constituted by mixing equivalent quantities of DNA from each individual.

Example 3 : Detection of the biallelic markers: amplification of genomic DNA by PCR

The amplification of specific genomic sequences of the DNA samples of example 2 was carried out on the pool of DNA obtained previously. In addition, 50 individual samples were similarly amplified.

PCR assays were performed using the following protocol:

Final volume	25 µl
DNA	2 ng/µl
$MgCl_2$	2 mM
dNTP (each)	200 µM
primer (each)	2.9 ng/µl
Ampli Taq Gold DNA polymerase	0.05 unit/µl
PCR buffer (10x = 0.1 M TrisHCl pH8.3 0.5M KCl)	1x

Each pair of first primers was designed using the sequence information of the *hGGPS* gene disclosed herein and the OSP software (Hillier & Green, 1991). This first pair of primers was about 20 nucleotides in length and had the sequences disclosed in Table 1 in the columns labeled PU and RP.

Table 1

	Amplified region of hGGPS gene	PU	RP
1	Partial Intron 3/Partial Exon 4	SEQ ID No 9	SEQ ID No 10

Preferably, the primers contained a common oligonucleotide tail upstream of the specific bases targeted for amplification which was useful for sequencing.

Primers PU contain the following additional PU 5' sequence : TGTAAAACGACGGCCAGT; primers RP contain the following RP 5' sequence : CAGGAAACAGCTATGACC.

The synthesis of these primers was performed following the phosphoramidite method, on a GENSET UFPS 24.1 synthesizer.

DNA amplification was performed on a Genius II thermocycler. After heating at 95°C for 10 min, 40 cycles were performed. Each cycle comprised: 30 sec at 95°C,

54°C for 1 min, and 30 sec at 72°C. For final elongation, 10 min at 72°C ended the amplification. The quantities of the amplification products obtained were determined on 96-well microtiter plates, using a fluorometer and Picogreen as intercalant agent (Molecular Probes).

Example 4 : Detection of the biallelic markers: sequencing of amplified genomic DNA and identification of polymorphisms.

The sequencing of the amplified DNA obtained in example 3 was carried out on ABI 377 sequencers. The sequences of the amplification products were determined using automated dideoxy terminator sequencing reactions with a dye terminator cycle sequencing protocol. The products of the sequencing reactions were run on sequencing gels and the sequences were determined using gel image analysis [ABI Prism DNA Sequencing Analysis software (2.1.2 version) and the above mentioned proprietary "Trace" basecaller].

The sequence data were further evaluated using the above mentioned polymorphism analysis software designed to detect the presence of biallelic markers among the pooled amplified fragments. The polymorphism search was based on the presence of superimposed peaks in the electrophoresis pattern resulting from different bases occurring at the same position as described previously.

Table 2 shows the biallelic marker that has been detected after the sequence analysis of the amplification fragments generated by PCR.

Table 2

Amplicon	Marker Name	Localization in *RBP-7* gene	Polymorphism	Major allele	Minor allele
1	5-187-77	Intron 3	A/T	SEQ ID No 7	SEQ ID No 8

Example 5 : Validation of the polymorphisms through microsequencing

The biallelic marker identified in example 4 was further confirmed through microsequencing. Microsequencing was carried out for each individual DNA sample described in Example 2.

Amplification from genomic DNA of individuals was performed by PCR as described above for the detection of the biallelic markers with the same set of PCR primers (Table 1).

The preferred primers used in microsequencing were about 20 nucleotides in length and hybridized just upstream of the considered polymorphic base. According to the invention, the primer used in microsequencing is detailed in Table 3.

Table 3

Marker Name	PU Microsequencing primer
5-187-77	SEQ ID No 11

The microsequencing reaction was performed as follows :

After purification of the amplification products, the microsequencing reaction mixture was prepared by adding, in a 20µl final volume: 10 pmol microsequencing oligonucleotide, 1 U Thermosequenase (Amersham E79000G), 1.25 µl Thermosequenase buffer (260 mM Tris HCl pH 9.5, 65 mM $MgCl_2$), and the two appropriate fluorescent ddNTPs (Perkin Elmer, Dye Terminator Set 401095) complementary to the nucleotides at the polymorphic site of each biallelic marker tested, following the manufacturer's recommendations. After 4 minutes at 94°C, 20 PCR cycles of 15 sec at 55°C, 5 sec at 72°C, and 10 sec at 94°C were carried out in a Tetrad PTC-225 thermocycler (MJ Research). The unincorporated dye terminators were then removed by ethanol precipitation. Samples were finally resuspended in formamide-EDTA loading buffer and heated for 2 min at 95°C before being loaded on a polyacrylamide sequencing gel. The data were collected by an ABI PRISM 377 DNA sequencer and processed using the GENESCAN software (Perkin Elmer).

Following gel analysis, data were automatically processed with software that allows the determination of the alleles of biallelic markers present in each amplified fragment.

The software evaluates such factors as whether the intensities of the signals resulting from the above microsequencing procedures are weak, normal, or saturated, or whether the signals are ambiguous. In addition, the software identifies significant peaks (according to shape and height criteria). Among the significant peaks, peaks corresponding to the targeted site are identified based on their position. When two significant peaks are detected for the same position, each sample is categorized classification as homozygous or heterozygous type based on the height ratio.

References

Barany F., 1991, Proc. Natl. Acad. Sci. USA, **88** : 189-193.

Beard et al., 1980, Virology, **75**:81

Beaucage et al., *Tetrahedron Lett* 1981, **22**: 1859-1862

Berthon P. et al., 1998, Am. J. Hum. Genet., **62** : 1416-1424.

Bruisten s. et al., 1993, AIDS Res. Hum. Retroviruses, **9** : 259-265.

Burg JL et al., 1996, Mol. and Cell. Probes, **10** : 257-271.

Chai H. et al., 1993, Biotechnol. Appl. Biochem., **18**:259-273

Chee et al., 1996, Science, **274**:610-614.

Chu B. et al., 1986, Nucleic Acids Research, **14** : 5591-5603.

Current Protocols in Molecular Biology, 1989, Ausubel FM et al. (eds), Greene Publishing Associates, Sections 9.10-9.14.

Duck P. et al., 1990, Biotechniques, **9** : 142-147.

Feldman and Steg, 1996, Medecine/Sciences, synthese, **12**:47-55

Felgner et al., 1987, Proc. Natl. Acad. Sci., **84**:7413

Flotte et al., 1992, Am. J. Respir. Cell Mol. Biol., **7** : 349-356.

Fraley et al., 1980, J. Biol. Chem., **255**:10431

Fuller S.A. et al., 1996, Immunology in Current Protocols in Molecular Biology, Ausubel et al. Eds, John Wiley & Sons, Inc., USA

Guatelli J C et al., Proc. Natl. Acad. Sci. USA, **35** : 273-286.

Guzman, 1981, Cell, **23** : 175

Hames BD and Higgins SJ, 1985, "Nucleic acid hybridization : a practical approach", Hames and Higgins Ed., IRL Press, Oxford.

Hillier L. and Green P. *Methods Appl.*, 1991, **1**: 124-8.

Houbenweyl, 1974, in Meuthode der Organischen Chemie, E. Wunsch Ed., Volume 15-I et 15-II, Thieme, Stuttgart.

Huang L et al., 1996, Cancer Res; 56(5):1137-1141.

Huygen et al., 1996, Nature Medicine, 2(8):893-898

Julan et al., 1992, J. Gen. Virol., **73** : 3251 – 3255.

Kaczorowski T, Szybalski W, *Anal Biochem* 1994;**221**(1):127-135

Kaneda et al., 1989, Science, **243**:375

Kieleczawa J, Dunn JJ, Studier FW, *Science* 1992;**258**(5089):1787-1791

Kievitis T. et al., 1991, J. Virol. Methods, **35** : 91-92.

Kohler G. and Milstein C., 1975, Nature, **256** : 495.

Kotler LE, Zevin-Sonkin D, Sobolev IA, Beskin AD, Ulanovsky LE, *Proc Natl Acad Sci U S A* 1993;**90**(9):4241-4245

Kwoh D Y et al., 1989, Proc. Natl. Acad. Sci. USA, **86** : 1173-1177.

Landergren U et al., 1988, Science, **241** : 1077-1080.

Leger OJ, et al., 1997, Hum Antibodies, **8**(1): 3-16

Lenhard T. et al., 1996, Gene, **169**:187-190

Levrero et al., 1991, Gene, **101**:195

Lin Z, Floros J, 1998, Biotechniques, **24**(6):937-940

Livak KJ and Hainer JW, 1994, Hum. Mutat., **3**(4) : 379-385.

Lizardi PM et al., 1988, Bio/Technology, **6** : 1197-1202.

Lockhart DJ, 1996, Nat Biotechnol, **14**(13):1675-1680

Mackey K, Steinkamp A, Chomczynski P, 1998, Mol Biotechnol, **9**(1):1-5

Martineau P, Jones P, Winter G, 1998, J Mol Biol, **280**(1):117-127

McLaughlin et al., 1989, J. Virol., **62** : 1963 – 1973.

Merrifield RB, 1965, Nature, **207**(996): 522-523.

Merrifield RB., 1965, Science, **150**(693): 178-185.

Midoux, 1993, Nucleic Acids Research, **21**:871-878

Miele EA et al., 1983, J. Mol. Biol., **171** : 281-295.

Muzyczka et al., 1992, Cuur. Topics in Micro. and Immunol., **158** : 97-129.

Narang SA, Hsiung HM, Brousseau R, *Methods Enzymol* 1979;**68**:90-98

Neda et al., 1991, J. Biol. Chem., **266** : 14143 – 14146.

Nyren P. et al., 1993, Anal. Biochem., **208**(1) : 171-175.

O'Reilly et al., 1992, Baculovirus expression vectors : a Laboratory Manual. W.H. Freeman and Co., New York

Ohno et al., 1994, Sciences, **265**:781-784

Ovyn C. et al., 1996, Mol. Cell. Probes, **10** : 319-324.

Pagano et al., 1967, J. Virol., **1**:891

Pastore, 1994, Circulation, **90**:I-517

PCR Methods and Applications" (1991, Cold Spring Harbor Laboratory Press

Pietu G, 1996, Genome Res, **6**(6):492-503

Reimann KA, et al., 1997, AIDS Res Hum Retroviruses. **13**(11): 933-943

Ridder R, Schmitz R, Legay F, Gram H, 1995, Biotechnology (N Y), **13**(3):255-260

Roth J.A. et al., 1996, Nature Medicine, **2**(9):985-991

Roux et al., 1989, Proc. Natl Acad. Sci. USA, **86** : 9079 – 9083.

Saiki R K et al., 1985, Science, **230** : 1350-1354.

Sambrook, J. Fritsch, E. F., and T. Maniatis. 1989. Molecular cloning: a laboratory manual. 2ed. Cold Spring Harbor Laboratory, Cold spring Harbor, New York.

Samson M et al., 1996, Nature, **382**(6593):722-725.

Samulski et al., 1989, J. Virol., **63** : 3822-3828.

Sanchez-Pescador R., 1988, J. Clin. Microbiol., **26**(10):1934-1938

Sczakiel G. et al., 1995, Trends Microbiol., 1995, **3**(6):213-217

Segev D. et al., 1992, *Amplification of nucleic acid sequences by the "Repair Chain Reaction"*. In : *Non-radioactive labeling and detection of biomolecules*. Kessler C. Springer-Verlag, Berlin, New York, pp. 142-147.

Shena et al., 1995, Science, **270** : 467-470.

Shena et al., 1996, Proc. Natl. Acad. Sci USA, **93** : 10614-10619.

Smith et al., 1983, Mol. Cell. Biol., **3**:2156-2165.

Sosnowski RG, Tu E, Butler WF, O'Connell JP, Heller MJ, *Proc Natl Acad Sci U S A* 1997;**94**(4):1119-1123

Spargo C A et al., 1996, Mol. Cell. Probes, **10** : 247-256.

Stone BB et al., 1996, Mol. Cell. Probes, **10** : 359-370.

Tacson et al., 1996, Nature Medicine, **2**(8):888-892.

Urdea M.S., 1988, Nucleic Acids Research, **11**: 4937-4957

Urdea MS et al., 1991, Nucleic Acids Symp Ser., **24**: 197-200.

Vaughan TJ, et al., 1996, Nat Biotechnol. **14**(3): 309-314

Vlasak R. et al., 1983, Eur. J. Biochem., **135**:123-126

Wabiko et al., 1986, DNA, **5**(4):305-314

Walker G T et al., 1992, Nucleic Acids research, **20** : 1691-1696.

White, M.B. et al., *Genomics* 1997; **12**:301-306.

What is claimed is :

1. A purified or isolated nucleic acid encoding a human geranylgeranyl pyrophate synthetase (*hGGPS*), wherein said nucleic acid comprises the nucleotide sequence of SEQ ID No 1 or a polynucleotide having at least a 95% nucleotide identity with SEQ ID No 1.

2. A purified or isolated nucleic acid encoding a human geranylgeranyl pyrophosphate synthetase, wherein said nucleic acid comprises a nucleotide sequence selected from the group consisting of :

a) the nucleic acids of SEQ ID No 4 and 5, or a polynucleotide having at least a 95% nucleotide identity with any of the nucleotide sequences of SEQ ID Nos 4 and 5;

b) a nucleic acid fragment of a nucleotide sequence selected from the group consisting of SEQ ID Nos 4 and 5, wherein this nucleic acid fragment encodes a polypeptide having an amino acid sequence beginning at the amino acid in position 200 and ending at the amino acid in position 300 of the hGGPS polypeptide of SEQ ID No 6, or a nucleic acid encoding a peptide fragment thereof.

3. A purified or isolated nucleic acid comprising a polynucleotide which is selected from the group consisting of the nucleotide sequences of SEQ ID Nos 2 and 3, or a biologically active fragment or variant thereof.

4. A purified or isolated nucleic acid of at least eight nucleotides in length, wherein said nucleic acid hybridizes under stringent hybridization conditions with a polynucleotide selected from the group consisting of the nucleotide sequences of SEQ ID Nos 2 and 3, or a sequence complementary thereto.

5. A purified or isolated nucleic acid comprising :

a) a nucleic acid comprising a regulatory polynucleotide of SEQ ID No 2 or a biologically active fragment or variant thereof;

b) a polynucleotide encoding a desired polypeptide or nucleic acid operably linked to the polynucleotide of SEQ ID No 2 or its biologically active fragment or variant thereof;

c) optionally, a nucleic acid comprising a regulatory polynucleotide of SEQ ID Nos 3 or a biologically active fragment or variant thereof.

6. A purified or isolated nucleic acid encoding a human geranylgeranyl pyrophosphate synthetase comprising the polynucleotide beginning at the nucleotide in

position 85 and ending at the nucleotide in position 987 of the nucleotide sequence of SEQ ID No 4.

7. A purified or isolated oligonucleotide useful as an amplification primer or as a probe, wherein this oligonucleotide is selected from the group consisting of the nucleotide sequences of SEQ ID Nos 9-11.

8. A purified or isolated oligonucleotide useful as an amplification primer or as a probe, wherein this oligonucleotide is selected form the group of nucleotide sequences consisting of :

a) A purified or isolated oligonucleotide beginning at the nucleotide in position 7233 and ending a the nucleotide in position 7251 of the nucleotide sequence of SEQ ID No 1;

b) A purified or isolated oligonucleotide which is complementary to the sequence beginning at the nucleotide in position 7565 and ending at the nucleotide in position 7582 of the nucleotide sequence of SEQ ID No 1;

c) A purified or isolated oligonucleotide beginning at the nucleotide in position 13582 and ending a the nucleotide in position 13600 of the nucleotide sequence of SEQ ID No 1;

d) A purified or isolated oligonucleotide which is complementary to the sequence beginning at the nucleotide in position 13982 and ending at the nucleotide in position 14001 of the nucleotide sequence of SEQ ID No 1;

e) A purified or isolated oligonucleotide beginning at the nucleotide in position 14222 and ending a the nucleotide in position 14240 of the nucleotide sequence of SEQ ID No 1;

f) A purified or isolated oligonucleotide which is complementary to the sequence beginning at the nucleotide in position 14626 and ending at the nucleotide in position 14645 of the nucleotide sequence of SEQ ID No 1;

g) A purified or isolated oligonucleotide beginning at the nucleotide in position 14606 and ending a the nucleotide in position 14623 of the nucleotide sequence of SEQ ID No 1;

h) A purified or isolated oligonucleotide which is complementary to the sequence beginning at the nucleotide in position 15007 and ending at the nucleotide in position 15026 of the nucleotide sequence of SEQ ID No 1;

i) A purified or isolated oligonucleotide beginning at the nucleotide in position 14845 and ending a the nucleotide in position 14864 of the nucleotide sequence of SEQ ID No 1;

j) A purified or isolated oligonucleotide which is complementary to the sequence beginning at the nucleotide in position 15246 and ending at the nucleotide in position 15265 of the nucleotide sequence of SEQ ID No 1.

9. A pair of oligonucleotide primers for amplifying a nucleotide sequence contained in the *hGGPS* gene, wherein said pair of primers is selected from the group consisting of :

a) : (1) Forward primer beginning at the nucleotide in position 7233 and ending a the nucleotide in position 7251 of the nucleotide sequence of SEQ ID No 1; (2) reverse primer which is complementary to the sequence beginning at the nucleotide in position 7565 and ending at the nucleotide in position 7582 of the nucleotide sequence of SEQ ID No 1.

b) : (1) Forward primer beginning at the nucleotide in position 13582 and ending a the nucleotide in position 13600 of the nucleotide sequence of SEQ ID No 1; (2) reverse primer which is complementary to the sequence beginning at the nucleotide in position 13982 and ending at the nucleotide in position 14001 of the nucleotide sequence of SEQ ID No 1.

c) : (1) Forward primer beginning at the nucleotide in position 14222 and ending a the nucleotide in position 14240 of the nucleotide sequence of SEQ ID No 1; (2) reverse primer which is complementary to the sequence beginning at the nucleotide in position 14626 and ending at the nucleotide in position 14645 of the nucleotide sequence of SEQ ID No 1.

d) : (1) Forward primer beginning at the nucleotide in position 14606 and ending a the nucleotide in position 14623 of the nucleotide sequence of SEQ ID No 1; (2) reverse primer which is complementary to the sequence beginning at the nucleotide in position 15007 and ending at the nucleotide in position 15026 of the nucleotide sequence of SEQ ID No 1.

e) : (1) Forward primer beginning at the nucleotide in position 14845 and ending a the nucleotide in position 14864 of the nucleotide sequence of SEQ ID No 1; (2) reverse primer which is complementary to the sequence beginning at the nucleotide in position 15246 and ending at the nucleotide in position 15265 of the nucleotide sequence of SEQ ID No 1.

10. A purified or isolated biallelic marker, wherein said biallelic marker is from the sequence of the *hGGPS* gene.

11. A nucleotide sequence comprising a purified or isolated biallelic marker according to claim 10.

12. A purified or isolated nucleic acid comprising a nucleotide sequence selected from the group consisting of SEQ ID Nos 7-8 or a variant or a fragment thereof, said fragment comprising at least 8 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID Nos 7-8 and including the polymorphic base thereof.

13. A method for the identification and characterization of a bialleleic marker in the genomic region harboring the *hGGPS* gene, said method comprising :

- providing a plurality of primer sequences capable of amplifying portions of the genomic region containing the *hGGPS* gene, and in particular portions of the polynucleotide of SEQ ID No 1;

- amplifying portions of the genomic region containing the *hGGPS* gene from a plurality of individuals using said primers to obtain a plurality of amplicons; and

- sequencing said plurality of amplicons to identify biallelic markers in the genomic region harboring the *hGGPS* gene.

14. A method for the amplification of the *hGGPS* gene or a fragment or a variant thereof in a test sample, said method comprising the steps of :

a) contacting a test sample suspected of containing the targeted *hGGPS* gene sequence or portion thereof with amplification reaction reagents comprising a pair of amplification primers located on either side of the *hGGPS* region to be amplified, and

b) detecting the amplification products.

15. The method according to claim 14, wherein the amplification primers are selected from the group consisting of SEQ ID Nos 9-10.

16. The method according to claim 14, wherein the amplification product is detected by hybridization with a labeled probe having a sequence which is complementary to a region of the *hGGPS* gene.

17. A kit for the amplification of a nucleotide sequence contained in the *hGGPS* gene, wherein said kit comprises :

a) A pair of oligonucleotide primers located on either side of the *hGGPS* region to be amplified;

b) Optionally, the reagents necessary for performing the amplification reaction.

18. A method for detecting the presence of a nucleic acid comprising at least a part of a nucleotide sequence selected from the group consisting of SEQ ID Nos 2, 3 and 7-11 in a sample, said method comprising the following steps of :

a) bringing into contact a nucleic acid probe or a plurality of nucleic acid probes, which can hybridize to a nucleotide sequence included in one of the nucleic acids of SEQ ID Nos 2, 3 and 7-11, and the sample to be assayed.

b) detecting the hybrid complex formed between the probe and a nucleic acid in the sample.

19. A kit for detecting the presence of a nucleic acid comprising at least a part of a nucleotide sequence selected from the group consisting of SEQ ID Nos 2, 3 and 7-11 in a sample, said kit comprising :

a) a nucleic acid probe or a plurality of nucleic acid probes, which can hybridize to a nucleotide sequence included in one of the nucleic acids of SEQ ID Nos 2, 3 and 7-11;

b) optionally, the reagents necessary for performing the hybridization reaction.

20. A recombinant expression vector comprising a nucleic acid selected from the group consisting of SEQ ID Nos 2 and 3, or biologically active fragments or variants thereof.

21. A recombinant expression vector containing a nucleic acid comprising :

a) a nucleic acid comprising a regulatory polynucleotide of SEQ ID No 2 or a biologically active fragment or variant thereof;

b) a polynucleotide encoding a desired polypeptide or nucleic acid operably linked to the polynucleotide of SEQ ID No 2 or its biologically active fragment or variant;

c) optionally, a nucleic acid comprising a regulatory polynucleotide of SEQ ID No 3 or a biologically active fragment or variant thereof.

22. A recombinant vector comprising a nucleic acid selected from the group consisting of :

a) a nucleotide sequence selected from the group consisting of : SEQ ID Nos 1, 4 and 5 or a nucleic acid having at least 95% nucleotide identity with a polynucleotide selected from the group consisting of the nucleotide sequences of SEQ ID Nos 1, 4 and 5;

b) a purified or isolated nucleic acid comprising a nucleic acid fragment of a nucleotide sequence selected from the group consisting of SEQ ID Nos 4

and 5, wherein this nucleic acid fragment encodes a polypeptide having an amino acid sequence beginning at the amino acid in position 200 and ending at the amino acid in position 300 of the hGGPS polypeptide of SEQ ID No 6, or a nucleic acid encoding a peptide fragment thereof.

23. A recombinant vector containing a nucleic acid comprising the nucleotide sequence beginning at the nucleotide in position 85 and ending in position 987 of the polynucleotide of SEQ ID No 4.

24. A recombinant cell host comprising a recombinant vector according to anyone of claims 20 to 23.

25. A recombinant cell host comprising a purified or isolated nucleic acid according to anyone of claims 1-3, 5-6 and 12.

26. The hGGPS polypeptide of the amino acid sequence of SEQ ID No 6.

27. A polyclonal or a monoclonal antibody specifically directed against a polypeptide selected from the group consisting of :

a) the hGGPS polypeptide of the amino acid sequence of SEQ ID No 6;

b) a polypeptide consisting in the amino acid sequence beginning at the amino acid in position 200 and ending at the amino acid in position 300 of the polypeptide of SEQA ID No 6, or a peptide fragment thereof.

28. A method for the screening of a candidate substance or molecule modulating the expression of the *hGGPS* gene, said method comprising the following steps :

a) providing a recombinant host cell expressing a nucleic acid, wherein said nucleic acid comprises a nucleotide sequence selected from the group consisting of SEQ ID Nos 1, 4 and 5;

b) obtaining a candidate substance, and

c) determining the ability of the candidate substance to modulate the expression levels of the nucleotide sequence selected from the group consisting of SEQ ID Nos 1, 4 and 5.

29. A kit for the screening of a candidate substance or molecule modulating the expression of the *hGGPS* gene, wherein said kit comprises a recombinant vector that allows the expression of a nucleotide sequence selected from the group consisting of SEQ ID Nos : 1, 4 and 5 or alternatively a recombinant host cell containing such a recombinant vector.

30. A method for the screening of a candidate substance or molecule modulating the expression of the *hGGPS* gene, said method comprising the following steps :

a) providing a recombinant host cell containing a nucleic acid, wherein said nucleic acid comprises a nucleotide sequence of SEQ ID No 2 or a biologically active fragment or variant thereof operably linked to a polynucleotide encoding a detectable protein;

b) obtaining a candidate substance, and

c) determining the ability of the candidate substance to modulate the expression levels of the polynucleotide encoding the detectable protein.

31. A kit for the screening of a candidate substance or molecule modulating the expression of the *hGGPS* gene, wherein said kit comprises a recombinant vector containing a polynucleotide encoding a detectable protein under the control of a nucleotide sequence of SEQ ID No 2 or a biologically active fragment or variant thereof.

(1) GENERAL INFORMATION:

(i) APPLICANT: Bougueleret, Lydie

(ii) TITLE OF INVENTION: A nucleic acid encoding a geranyl-geranyl-pyrophosphate synthetase (GGPPS) and polymorphic markers associated with said nucleic acid.

(iii) NUMBER OF SEQUENCES: 11

(iv) CORRESPONDANCE ADDRESS:
(A) ADDRESSEE: Knobbe, Martens, Olson & Bear
(B) STREET: 501 West Broadway
(C) CITY: San Diego
(D) STATE OR PROVINCE: California
(E) COUNTRY: USA
(F) ZIP: 92101-3505

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy Disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: Win95
(D) SOFTWARE: Word

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: Israelsen, Ned A.
(B) REGISTRATION NUMBER: 29,655
(C) REFERENCE/DOCKET NUMBER: GENSET.034PR

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: (619) 235-8550
(B) TELEFAX: (619) 235-0176

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17131 base pairs
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: DOUBLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: DNA

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

(ix) FEATURE:
(A) NAME/KEY: Polymorphic fragment 5-187-77 SEQ ID7
(B) LOCATION: 14036..14081

(ix) FEATURE:
(A) NAME/KEY: Polymorphic fragment 5-187-77 SEQ ID8
(B) LOCATION: 14036..14081

(ix) FEATURE:
(A) NAME/KEY: ex1
(B) LOCATION: 486..546

(ix) FEATURE:
(A) NAME/KEY: ex1bis
(B) LOCATION: 633..826

(ix) FEATURE:
(A) NAME/KEY: ex2
(B) LOCATION: 7292..7384

(ix) FEATURE:
(A) NAME/KEY: ex3
(B) LOCATION: 13760..13830

(ix) FEATURE:
(A) NAME/KEY: ex4
(B) LOCATION: 14063..15251

(ix) FEATURE:
(A) NAME/KEY: start CDS ATG
(B) LOCATION: 7315..7317

(ix) FEATURE:
(A) NAME/KEY: stop CDS
(B) LOCATION: 14822..14824

(ix) FEATURE:
(A) NAME/KEY: polyadenylation site
(B) LOCATION: 15126..15131

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:AA398854
(B) LOCATION: 486..546

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:AA398854
(B) LOCATION: 7292..7384

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:AA398854
(B) LOCATION: 13760..13830

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:AA398854
(B) LOCATION: 14063..14314

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:Z44596
(B) LOCATION: 633..826

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:Z44596
(B) LOCATION: 7292..7384

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:Z44596
(B) LOCATION: 13760..13830

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:AA435858
(B) LOCATION: 14243..14670

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:AA194600
(B) LOCATION: 15055..15251

(ix) FEATURE:
(A) NAME/KEY: upstream amplification primer 5-185
(B) LOCATION: 7233..7251

(ix) FEATURE:
(A) NAME/KEY: downstream amplification primer 5-185
(B) LOCATION: complement 7565..7582

(ix) FEATURE:
(A) NAME/KEY: upstream amplification primer 5-186
(B) LOCATION: 13582..13600

(ix) FEATURE:
(A) NAME/KEY: downstream amplification primer 5-186
(B) LOCATION: complement 13982..14001

(ix) FEATURE:
(A) NAME/KEY: upstream amplification primer 5-188
(B) LOCATION: 14222..14240

(ix) FEATURE:
(A) NAME/KEY: downstream amplification primer 5-188
(B) LOCATION: complement 14626..14645

(ix) FEATURE:
(A) NAME/KEY: upstream amplification primer 5-189
(B) LOCATION: 14606..14623

(ix) FEATURE:
(A) NAME/KEY: downstream amplification primer 5-189
(B) LOCATION: complement 15007..15026

(ix) FEATURE:
(A) NAME/KEY: upstream amplification primer 5-190
(B) LOCATION: 14845..14864

(ix) FEATURE:
(A) NAME/KEY: downstream amplification primer 5-190
(B) LOCATION: complement 15246..15265

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

TCGGGCTCCC TGGTTGGGGG GAGGGGGACG ACGAAAAATC CCCCCCGGAC TGGAGGTCCG 60

GGCCCCCAAT CGCGCTGCCC TCCAGAGGAC GGCGGCGATG GACCCTCTGC AGCTCCCTCC 120

GGGCAAAGGT CCAGGCGGTG GCCGTGGCGG CGGCAAGATG AAGCTCAAGA GTCTCCCTCC 180

GCTTCGGCGA CCGAGCTCCT CACTCCGGAC TCGACTGACG GGCAAACATC GCTTCCCCCC 240

CACCGACTCT AGGTTCCCCC CCTTTCTCCC CTCCCCTAGA TTTTTTTTCC CCCCCTCCCC 300

TACCTCTTTC CCGGATGGCC TCTTAGACGA CCTTGGATTG GTTAAAGTTC TTTAGAACCC 360

GCCTATACAC TGTTCCTATT GGTCCCTGGA TACAAACAAC GACGCCATTT TCCCACCAGT 420

TCTATGGAAA CAGAAAGTTA CGCCTCAAGG CTTTCTGGGA AATAAAGTCC ATACTCTGGG 480

GCCAACGCGC AAATCCTCGT CCGCGAGAAC TGCAAGGCCC GCAATGCCCT GCGCCTGCGT 540

GGACCGGTGC GGGGGCGGGG GGGAGGTGAA AGGGGCGGGG CAACAAAGCA GTAGGGAGGC 600

GGCAACGACG CCTGCGCAGT GTGACCGGGA TGGCGCATTT TCTTGCACCA ACTAATGCGG 660

TGTCGCTGGC GGCTGAGGAG GGCGGAGAGT TCTGTGGTGA AATAGTGGGA AGGATTCATG 720

TAGGCATCGG GAAGAGCCTA AGTCCACATT ATAAAATAGG AAGTTGATGC GGGGTACAGT 780

TACTCCCGGA CCGGCGGCGT GAAAGTCGTG ATATCATCGT TGAACTGTGA GCGGCAGTGG 840

CGGCGGCTGG GGGGAACCCG GATGGGAAGA AGGGCGGGGG AGGCTGGGAG GCGGGGCAGA 900

GGAAAGAAAG AAAGGAGAGT GAGGACCCGG ATGCTGAACC GGATTGTGTA TGAATTTTCC 960

ATCCCCTAGC TTTAAGCGAG GAGGGAGAGG AAGGGTTGGC CAAGTGGGGC GGAAGGGAGC 1020

ATCTGAGCGA GGAGGAAGCA GAAACCTCAC CGTTTCTTCC CCTCCGGACT CTGTGCTAGC 1080

ACTGTATACG TTTGCAGTTC TCTGCCCAGC CGCTGTGGAA AATCGGCCTC GAAGTGATTG 1140

AAATTCCCTG TTTATATCAG GCGGCTTCTT TCAGATCCAT CGTCTTTCTC CCGGAGTATG 1200

AATGGAAGGA TTCAGTATGC GCTTCACATT TGTATGTCTC TGGCCATTCT CAAACCAGGC 1260

CCTTCCCTTT GAAAAGTCTT TTGCATGGGA TGTTCACTTC TTAGACGCAA GGTTGTGTGC 1320

CCTGGTTTCA TCGTCTAACG CGTTAGAAGG CGCTTTCATT TCTTCATGGG TGTTGAGCGC 1380

CGACCACTGG GGTGGCCTCT GCCTTCGTAG ACCTGCGCCT GGTGAGACGG ACAGATGCTG 1440

AACAAAACGA TGTGAAATTA CCGCAGTGGC AGTGCCCCAG AGGAGAGTTC CACGGTGATA 1500

GGAGAATGAG GGAATTTGGC TTCTTTAGGG AGGGAAAGGA AGGGTTTCTG AGCAAGTGAG 1560

GATCGAGCTG AGAGCTGAAG GGCTAGCAGG AGTTAACTAA GGAAAGAGAA AAGGAAAAGA 1620

CATTCCAGAC AAAAAGGCTA ACTTGTCAGA AAGCCCTGTG GCGGAAGGGA GCTTTTCCAA 1680

TATGAAGAAC TGAGCCTGGA GAGATGGGAT GAGGGGGAGT GTCGAACCTT TTAGGCTTTG 1740

TAAAGGAGTT TTGGTTTTCT CCTAATAGCA ATGGGATATC TTCCAAGGAA TCTCAATCAA 1800

AAGGGAGAGA TGGCTCCGAT TGGAATGTCA TCCCTGGCTG AAGAGTNNAG GAAGCGAAAA 1860

AAAGAAGAGT TAAAGAGGCA AATGCAGGGA ACCCGACGAG GAGGCTATTG CCGTAGTAGT 1920

TCACATGGTG AAAAGAATGG AGCGTTTGTA TTAATGATTA TGGATTCACT CTTTGAACAA 1980

ATTTCTGGCA GCTTTTTAGT TTTGAAAGTG AGAAGTTTCA GACTCTCACT GAGGTATTCT 2040

GTAGTTTTTT CACTCTAAAA GGAAACTAGT AGAGTTCATG TAACACACAC TAATGCCTCT 2100

TTACATTTAA CTTTAGTATG TGATAGCTGA AATTTCCAGC TGTGATAAAT TGGGAAATCC 2160

TTTGATTTAA AAGAAAAACA AAGGCGGGTG AGGGTGAGAG TATATGCCAC GGTGTGTAGA 2220

ATCCTTTAGA CTCTTAAGAA GACACANGGC GGCTGGGCGT GGTGGCTCAC GCTTGTAATC 2280

CCAGCACTTT GGGAGGCCGA GGCGGGCGGA TCACGAGGTC AGGAGATCGA GACCATCCTG 2340

GCTAACACGG TGAAAGCCCG TCTCTACTAA AAATACAAAA AAATTAGCCG GGCAAGGTGG 2400

CGGGCGCCTG TAGTCCCAGC TACTCGGGAG GCTGAGGCAG GAGAATGGCG TGAACCCGGG 2460

AGGCGGAGTT TGCAGTGAGA CGAGATCACG CCACTGCACT CCAGCCTGGG CGACAGAGTG 2520

AGACGCTGTT TCAGAAGAAA GACACAAGGC AAGTTGGTTG TCGATACCTG GAAAAATTGA 2580

AGTTCTTATG TTTTCATACC ACTGAAAATG CTTGTATGTA AATATCCTCT GGGACAGGAA 2640

ATTGACTTAA GTGAGTATTC TTAAACATCT CTAAGTGAGG AAAGGAAATA TTTTTTAAAG 2700

CATAATTAGT GTTTTAAGTT GAAAAATAAC ATCAACCACA AAGCTCTACG AATTGAAACA 2760

AAGATTAGCT CTGATTTCTG TGCAACAGGG TACACCTGTT ACAGGTCCTG ACACAAAAGG 2820

GAATTCTGAA AGTGCATCTC ATTGATTTTT AAGTTCGGTC AAATGTGTTT TGGAGGCTGT 2880

GAGAAAATAT ACAAACGTGA TTCTTGCTCC CAACTTGTAG TTGAGAAAAG ATAGATACTA 2940

ACATTTAAAT AGAGAAGTAT ATGAGATCCT TTTTTAATTC TACTTTTAAT GATGTTCGAT 3000

AATAATCTTT TAGCTAAGCC ATTATTCTTC CTGTTTTGCA TCTTCTTTTC TTACTTCAAT 3060

CCCTGATAAT AAGGTCACGT GTCAGAGATC AAATAGTATA GGTAATAGGT TACCTAAATA 3120

GGTATTTGCA TAATAGGTTA CCTAACTAAA TAGGTTTTTG CCTAATAGGT ATGTTGATTA 3180

TTTCGCTTAC TTGATTCTTT ATGAGCCTTT TTTTCCTTGC GACGTCTTTG GTATTAATTG 3240

TTAGTCAAGA TGGATGTAGA AATTTTCCAT ATGGGATGTT TCTCTTTGAA TTCATGTTGT 3300

TAAAATGATT TCTTTTGGTG GAGTGCTGAT CTTTTTTATG ATTGTTTCAT ATAGATAAGA 3360

ACAGACTACA AAAAAATATG CCTTTCAATC CTGAAGAGTA ACCTGAACTA TACACTAGTT 3420

TTGTGCTTTA ATTTTCATTT GTAATCTGCC TTCAATAAAG AGTTAAGCTA GTGGAATTTA 3480

TGTCTTAGCT TGTTATAACA CAAACACGAA TATTTGTCTG CTTGGCATTA AAGGGTAAAG 3540

ATATTCCATA GCTGGGAATC TTAATCTGAG GTACGTGTAA ACATTCAGGG ACTATATGAT 3600

CTCTGAGAAT TTGTATGTTG TAAGTCTTTG TGGCAGTGTA TACATTTGTG TTGCAACTTA 3660

TTAACACATA CACCGGGCTT TTTTTTTTTT TTTTAGAAGA TTCATAGCTT TCATCATATT 3720

CTCAAAAGGT TTCTGTGACC CATGAGATGG TTTACAGTAT GGGGAAGCAT CAAAGCACTT 3780

GCACAGTTGA TGGTTATATG TGTGTGTTAT TATTTCAGCC ACCCATTATC ATGTGCTTAC 3840

CAACTGCCTA ACAGTGCATA CATATGTAGA AGTTTTATTC TTTTCTCCTG TTGCCATATT 3900

ATACGTCTCA TTTCACAGCA GAAAAACAAC TGCATGACAG AGACAATGTG GTTCAAACCA 3960

TTTTACCCTT GTATTCATTG ACTGCTACAA AACAGGAACA TTAAATACCT GATTGTCACC 4020

AAATTGGGTA GTCTCAGCAC TTCTACACTC GTAATTGTGC TGGAAAAGTG GAATGCTAGC 4080

ACTAATAATT AGATTTTGGT TTGGAGGGTT TTTTATTTGT TTATTCTTAC TTGTATAAAT 4140

TTATGGGGTG CAAGTGTAGT TTTATCACAT GCATAGATTG CATTGTAGTG AAGTCAGGAC 4200

TTTTAGGGGG TCCATCACCC ATGTAATCAC GTTGTACCCA TTAAGTAATC TTTCATCATC 4260

CACCTCCTTC CCACCTTCTC ACCCTTTGGA ATCTCCATTG TCTATCATTC CACACTCCAT 4320

GTCCATGTAT ACACATTATC TAGCTCCCAT TTATAATTGA GAAGATGTAC TATTTGTCTT 4380

TTATGTCTGA CTTGTTACAC TTAAGGTAAG GGCTATCCAT CCATTTTGCT GCAAATGACA 4440

TGATTTCATT TTGTTTTAAT GGCTGAGTAA TCATTCGTTG TATATATACC ACATTTTCTT 4500

TATTCAGTCA TCTGCTGATG GACACTTAGG TTGATTCCAT ATCTTTACTA TTGTGAATAG 4560

TGCTGTAATA AACACATAGT GCAAGATTTT GGAAATTTTA CTTTTGTGGC ACGTTGTTGG 4620

TATTTACTCA GGATCTTTGG ATTTGCTTGG CTGCATGTAT ATGAATCAGT GTGTTTATTT 4680

ACTGAAATAT GTGCAAAAGT CTTGTCTTTG GTGGATTAAT TTATAATATA AATCCACAAA 4740

AGTCAGATTC TGCTCCTAAG TATATTTTAC ATTTTTAAAT TTAATGCCAG CAAGAAGTTA 4800

CAGTACTAGA ATTGCCTTAC CCCTGAGAGT ATCAATGATC AGATCATAGT ATCAGGTGAC 4860

TGGGCTATAG AAGATGACTT TTATTACTTA ACATTATGAA GTTACTAGGG CTGATTTAGA 4920

AATCGAGGAA CACTGGTGAA ACCCCGTCTC TACTAAAATA CAAAAATTAG CTGGGCGTGG 4980

TGGTGGGCAC CTGTAGTCCC AGCTACTCAG AAGGCTGAGT CAGGAGAATT GCTTGAGCCC 5040

AGGAGGCAGA GGTTGCAGTG AGCCGAGATC GTGCCACTGC ACTCCAGCCT GGGCGACAGA 5100

GTGAGACTCC GTCTCAAAAA AAAAAAAAAA AAAAAAAAAG GAACACATCC TCACTGTTAC 5160

AATAAATAAC AGTAGCCCAC ACCCCCTTAG TTGTGATGTG GTGTGATACC ATGTAAGCAA 5220

CCTATTTCCA GTTCCCCTAA CATTCTCAAG CAGCTGTATC AGAATCATAC AAGATGCATA 5280

TTTAAATTGA AGATTTCTAA GTCTCTGGCC CAGACTTAGA AAAAAAGGAT CAGGCCGGGC 5340

ACAGTAGCTA ACACCTGCAA TTCCAACACT TTGGGAGGCT GAGGCGGGTG GATCGCCTGA 5400

GGTCAGGAGT TTTGAGACCA GCCTGGCCAA CATAGTGAAA CCCCATCTCT ACTAAAAATT 5460

CAAAAAATTA GCTGGGCGTG GTGGCAAGAA CCTGTAATCC CTGCTATTCG GGAGGCTGAG 5520

GCAGGGGAAT CACTTGAACC CGGGAGGTGG AGGTTGCAGT GAGCCAAGAT TGCGCCACTG 5580

CACTCCAGCC TGGGCAACGA GCAAAACTCC GTCTCAAAAA AAAAAAACAA AAGGACCTTT 5640

GAGCAATCAG AATAACACAA AGTACATGAA CTGAACTTCA TTTTCTTCAT TCAAAAGAAA 5700

GTGGCCCTCA CTCAAGCAAA TATATTCTTG TGCTTTATCT TCTGGCATAC TGAGATAACT 5760

TTCTAAAGTG GTTTCCAATT CCAAAATCCA ATGATGTGCA ACTCATTGAA CAGCCCTAAC 5820

CACAAACTGC CATTAGATGC CATATTACAT TTAGCCTTTT TGTTGTAGAA AAGTTGGTTA 5880

GAAGTGGGCT CAGGATTCTA AAGACTAAAT CATAGTCCCA AGAAGCAAAA GAAAGAGGAT 5940

AAAAGTAATA AACTTCCCAA AATGTGCCAA AGATGCTAGA GCAGTTAGAT TCCTAATATG 6000

AGGACAAGTA ATAATAGAAA CAGATACAAA GAAATAAAGT AGAGATTCAA CAGTACAGGG 6060

AGACCCTAGG AAGACCATGA GTGTTATTCT AGGAAATACT GAAATAAGAC AGATTTCAGT 6120

ATAAAGGGGN AATATGTTTA ATAANATATA TGCATTTGAG TTAATGCGTA TTTTAAATCA 6180

GAAATCTCTG AAATGGATTG ATTGTAGAGA AACTACTAGG GGGACGAGGA GAATCCCTTT 6240

AAATTTTAAA TACATAAAAC ATACTCATCT TAGTGCTCAT TTAAAAAAGG ATATGTTTAC 6300

TAATTAGTGT AATCAGTTAA ATACAGAGGT ATCTTTCCAA TTCTTTGGAT GTGTTTTGAC 6360

ATTTGCCGTC AACNAATTAA GCCTTTTGTG GTTGATTAAA ATAGGAAAAG CTTAATATAA 6420

GTTATGTGAC TAAGAAAACA ACTTAAAAAC CAAGACAACA CTTTGACCAA TATAATCACT 6480

TGAATGAAGA ATTTTCTAAT TGAGATATAA TTTACATACC ACCCATTTAA AGTGTACATT 6540

TCAGCAGTTT TTAGTGTATT CACAGGGCTG TGCAACCATC ACAATTTAAT TTTATAACAT 6600

TTTGATCCCT GCGAAAAGAA ACCCTGTACT CATTAGCAAT TAGTCCCTGT TCCTAACCAC 6660

TAATCTACTT TCTTTCTCTG TAGATTGGCT TATTCTGAAC ATTTCGTATA AATGGAATCA 6720

TACAATATGT AGTCTCTTGA GATTGGCTTC TTTCACTTAA CATGTTTTCA AGGCTTCATA 6780

GCTGTAGAAT CTTGCTTTGT TTTTTTGAGA CTGGAGTCAC TCTTTCGCCC AGGCTGGAGT 6840

GCAGTGGTGT GATCTCAGCT CACTGCAACC TCTGCCTCCC GGGTTCAAGC AGTTCTCCTG 6900

CCTCAGCCTC CCAAGTAGCC AGAACTACAG GCACACACCA CCATGCTCGG CTAATCTTTG 6960

TAGTTTTAGT AGAGATGGTG TGAAGGCTGG TCTCGAACTC CTGACCTCAT GATCTACCCA 7020

CCTCAGCTAA TTTTTCATAT TTTTAGTAGA GACAAGGTTT TGCCATGTTG CCCAGGCTGG 7080

TCTCGAACTC CTGGGCTTAA GCTATCCGCC CGCCTCAGCC TCCCAAAGTG CTGGGATTAC 7140

AGGCGTGAAC TACCGTGCCC AGCAACAGAA TCTTCTTTTT AAACCAGACT AGGTGTCTTT 7200

TCACAAACAC CCTGCAATAC AAATTCCTTT GCAGTTTGAC ACTGAAAGAT GATTAGTTTC 7260

ATGTGATCTT TATGTTTCTC CTTTTTGACA GATTAGCTTT GAAGTTTAAA TCCAATGGAG 7320

AAGACTCAAG AAACAGTCCA AAGAATTCTT CTAGAACCCT ATAAATACTT ACTTCAGTTA 7380

CCAGGTAATA CTTCACTTAC AGTCCATATA GGGTCATTTT CATGCAGTAG TGGTCGTTCA 7440

AATGTTAGCA AATAGAAAAG GTTAGACTTG CTAGCCGTTG AGATTTTCTA TTTAAGGTGA 7500

TGCGTATGAG AAAAATGATA AATAGAACAT TATAATTTTT TCTTTATTAA AAGGTAATTT 7560

TTGCCAGGTG CAGTGATACA TACCTGTTGT CCCACCTACT TGGGAGGCTG AGGCAGGAGG 7620

ATGGCTTGAG CCCAGGAGTT TAAGGCTATA GTGCACAATG ATCACACCTG TGAATAGCCA 7680

CTACACTCCA GCTTGGGCAA CATAGTGAGA CCCCGTCTCT TAAAAAGAAA CGTAATTTTT 7740

GAAGGCACCC TTTAAAACAT ATCCAATTAT TTAACATATC TTGAAAAATA AAAATACTTA 7800

AAACATTTTG GTATCTCATT GGAGGTTGTA CTCTTTACGG ATATTACGCA TTCAGATTCC 7860

CCACTGTTTA NATATTAGGG GAAGTTACGC AGATTTGTTT AACAGTAGAA CACTTTATTT 7920

ACCATACATG TTCAAGTTTA CCTTCTATGT CTGTATTTTC CAGTATCTCA CACATACACT 7980

GCATTTCATA TACTACTGGT TCCTTTGAGA GCCAAATAAT AATGTATCTA AAATCACAGT 8040

ATTTGGAAAT ATAGCCCACT TTATTCCTGT ATAAGGGTAT GCCACCTTGG ACATGGCTTC 8100

CTACCTCACG TGTACGTGTG TGTTTTTGTT TTATTTTGCT TCTTTAAAAA CTTGTCTGGA 8160

GGCTGGGCGT GGTGGCTCAC GCCTGTAATC CCAGCACTTT TTGAGGCCAA GGCGGGCGGA 8220

TCATGAGGTC AAGAGGTTGA GACCAGCCTG GCCAACATGG TAAAACCCCG TCTCTACTAA 8280

AAACACNAAA GTTAGCTGGG CATGGTGGCG CATGCCTGTA GTCCCAGCTA CTCGGGAGGC 8340

TGAGGCAGGA GAATCACCTG AACCTGGAAG GCAGAGGTTG CAGTGAGCTG AGATTGCATC 8400

ACTGCACTCC AGCCTGGCAA CAGAATGAGA CTCCGACTCA AAAAAAAAGA AGAACTTGTC 8460

TGGAAATGAT AATAAGCAAA AACTCATGAA TATAATAAAC AGGGGTTATT GTAATAAAAA 8520

ATCATTTGTA TTAGAATATT CTTTCTCATA GACATAATAT AGGCCAGGTG TGGTGGCCCA 8580

CACCTGTATT CCCAGCACTT TGGGAAGCCA AGGCAGGATT GCTTGAGACC AAAAGTTTGA 8640

GACCACCTTG GGCAACATAA CAAGTCCCCC TCTCTGTTTT AAACATTTTT TAAAAAAGAA 8700

GAAATAATAT AAAAGTTGGT AAATTATTTG ACAAGCATAA AAACCTATTT AGCCATACTG 8760

TGACTAAACT CTAATGATGC TCTCAATTCA GTCTCAATAG ACACTTTTAA ATTTCCGTGC 8820

TAAAGTACAC ACCTTTCTTT ATGAGCACTT CTCTGTGGTA ATATGTGCAT TTCTGTTCTT 8880

CATGAGCCTG GGAAGGATAA AAGCCAAAAG AATGCTTGCT CCTGTGCTAC ACCTTGGAAA 8940

CCATAATTAG TGTCATTTTT ATTTTGGCCG ACCCTAATAG AGACTCGCCT GCTAATGTCA 9000

ATGCATGAGA AGAATGAGGG AATGACAGAA ATGGAGAATT CAAAGGGAAG GTTGCCCACT 9060

GTTTAAGAAA AAGCCAAGAG ACTGCTTTTG AGTGACATTT ATCCAGCAGT TAGTAACTTA 9120

TTTCAGTATC TCCCAGTGAG AAACATGGCA CAGTTTCACT TTCACTCTAC CCAGCTCTTA 9180

CTGCCAGACA TCCTTTAGAA CACGCTCACA AACACTAGCT GGAACTGGGC TGGCATTAAT 9240

AGCAAGCCAG TTATCAGTGC TGACAAAAGT CTAACAAGCA TCGCTTGAAT GTCTCTTACT 9300

CTGCTACTTA CAAAGCAAGG ACTGCCTACA GTTACATTTT AACCATAATG CTTACTTATG 9360

CTGTGACCAC CTTCTGTGAC TTCCTTTTTT TTAATTCTCA TTACTTGGAA ATAATGTTTT 9420

AAGACATTAG ATAACATATT TAAAATTATC ACTAGGTACC TCACCTTTTT ATTCAAGTAC 9480

GTTCTTGATC CATGATGGAA TACAACCTCA AAAGATACTA CTAAAGAAAT ATGACATTGC 9540

ACTATGCACA TAACACACTT ATTTTTTTAC AGAGAGCTTC AGAGTTACTA AAGTAACTTA 9600

GAGGTGTGCC AGGTCATTTA TACTGTTGTA ATATTACTCT TGCTAATAAA TAATAATAAT 9660

GCTATCAGTA TTTTCTGAAG TCAACCTGGC CAACATGGTG AAACCCTGTA TCTACTAAAA 9720

ATACAAATAT TAGCCAAGTA TGGTAGCGCA TGCCTGTAGT CCCAGCTGAG GCACGGGAGT 9780

CACAGGAGCC TAGGAGGCAG AGGTTGCAGT GAGCCGAGAT CACGCCACTG CACTCCAGCC 9840

TGGGCAACAG AGTGAGACAC TGTCTCAAAA AAAAAAAAGG ATTTTCTGAA ATTAGTAAAG 9900

AAAATTATTT TTATTTTTAA ATTTCTCATA CTTGCTGTCA TCTTATGTTT ATGTTTGTTT 9960

ATTTGCCTTA GTGTGGGGCC CTAGATGAGG TGAAGGGTGG GATTAGGGAG AGATGAAGCT 10020

GGCAGTGGAG GAAGAAGGGC TCCAAAAAGA GAGACAATAA TGTTTAGATC TTAAAGAGGA 10080

AGCAGTAATC TTTTAATTTT GAGAGATCTC TGTGATTAGC CTCAGTACTA GAAATTATTT 10140

TGGAACTCAG CCAGGCGCGG TGGCTCACAT CTGTACTCCC AGCACTTTGG GAGACCGAAG 10200

TGGGCAGATG GCTTAAGCCC AGGAGTTCAA GACCAGCCTG GGCAACATGG CAAAACCCTG 10260

TCTCTACTAA AAATACAAAA AATTAGCCAG GCATGTGATA CGCCCTTGTA GTCCCAGCTT 10320

ACCTGGGGGA CTGAGGTGGG ATGATTACCG GAGCCTGGGA GGTTGAGGCT GCAGTAAGCC 10380

AAGATCACAC CACTGCACCC CAGCCTGGGT GATTAAGGGA GACCCCGTCT CAGAAAAAAA 10440

AAAGGGGGGG AAACTTAAAA GCATCAGGCT AAACACTAGC ATGTCATCAG AGGGGAAAAA 10500

AATATTAAAA CTGTAGTACC TCAAAAATAA GCCATATATT GTACTGTTTT CTATATAACA 10560

TTCAAAAGTA AAATGAAAAA TGAAATTTCA CATTGAGACT CTGTTTTTCA TCTTCAAAAA 10620

AATGTGTTTA AGTGATACAG GCCAAGTGCA GTGGCTGACT TATTATCCCA GCACTTTGGG 10680

AGGCCAAGTG GGACAGATTG CTTTTGAGCC CAGGGGTTTG AGACCAGCCT GGGCAACAGG 10740

GCGAAACCCT GCCTCTACAA AAAATAAATA AATAAAAATA AAATTAGCCA GGCATGGTGG 10800

CTTGTTCTTG TAGNTCCCAG CTACTCAGGG GACTTGAGCC TAGGAGGTCA AGGCTGCAGT 10860

AGGCCGTGAT TGTGCCACTG CACTCCAGCC TGGGTGACAG AGCGAGACCC TGTCTCAAAA 10920

ATAATAATAA TAGGCCGGGC GTGGTGGGTC ACACCTGTAA TCCCAGCACT TCGAGAGGCC 10980

AAAGCATGTG GACGACTTGA GGTCAGGAGT TCGAGACCAG CCTGGCCAAC ATGGGGAAAC 11040

CCTGTCTCTA TTAAAAGTAC AAAAAATTGG CCGGGCGCGG TAGCTCACGC ATGTAATCCC 11100

TACACTTTGG GAGGCTGAGG TGGGTGGATC ACCTGAGGTC AGGAATTCAA GACCAGCCTG 11160

GCCAACATGA TGAAACCGTC TCTACTAAAA ATACAAAAAA TTAGCTGGAT TTAGTGGCGC 11220

ACGACTGTAA TCCCAGCTAC TCAGGAGGCT GAGGCAGGAG AATCGCTTGA ACCTAGGAGG 11280

TGGAGGTTGC AGTGAGCCAA GATCGTGACA CTGTACCCCA GCCTGGGCAA CAAGAGCAAA 11340

ACTCGATCTC AGAAAAAAAA TACAAAAAAT TAGCTAGGCG TAGTGACGCA CACCTGTAAT 11400

CCCAGCTACT CGGGAGGCTG AGACAGGAGA ATCCCTTGAA CCCAGGAGGC GAAGGTTGTG 11460

GTGAGCCGAG CCAAGATCGT GCCATTGCTT TCCAGCCTAG GTGACAGAGC AAAACTTCAT 11520

CTCCACAAAC AAACAAACAA ACAAAAAAAC CCATAATCCC AGCATTTTGG GAGGCCAACA 11580

CAGGTGAATT ACCTGAGGTC AGGAGTTTGA CACCAGCCTG GCCAACATAG TGAAACCCTG 11640

TCTCTACTAA AATTACAAAA ATTAGCCAGG TGTGGTGGCA GGTGCCTGTA ATCCCAGCTA 11700

CTTGGGAGGC TGAGGCAGGA GAATCGCTTG AACCCAGGGG GCGGAGGTTG CAGTGAGCCG 11760

AGATCACACC ATTGCACTCT AGCCTGGGTG ACAAGAGCGA AATTCCATCT CCAAAAAAAA 11820

AAAAAGAAAA CAGTATTTTA GTTTTAACTT TTTATGTAAC CATTTTCCTG AAACCTTATC 11880

TAAAATTAGG ATGTTATTAC CATGCATTCA TTTAGCAGAA AACTTATAGA ACATTTTTAC 11940

TAAGTGAACT GGCCATGGTT TTTATCTATC ATTCCTTTGT ATGTGACTAC AATGACTTCT 12000

AGTGGTAACT TCTATCCAAA GACCTATCTT AAATTAGCCA GGCATGGTGG CACATGCGTG 12060

TAATCCCAGC TACTCAGGAG GCTGAGGCAG GAGAATAGCT TGATCTTGGG AGGCGGAGGT 12120

TGCAAGTGAG CCGAGATCAC GCCGCTGCAA TCCAGCCTGG GCAACAGAAT GAGACTCCGT 12180

CTCAAAAACA AAAAACAAAA AGACCTATCT TGAGCTTTCC GTGTAAGAAA AAGATGATAC 12240

TGTTGGGTGA AGTGACTCAA CGTCTGTAAT TTCAGCAATT TGGGAGGCTG TAGCGGCCGG 12300

ATTGCTTGAG CCCAGGAGTT TGAGACCAGC TTGGGCAACA TGGGAACACA CTGTCTCTAC 12360

AAAAACAAAA ATTAACCGGG CGTGGTCGCT TGCACCTATA GTGCCAGCTA CTCGGGAGGC 12420

TGAGGTGGAG GCTGCAGTGA GCTGTGAACA CACCACTGCA CTCCAGCCTG GGTGACAGAG 12480

TGAGACCCTG TCTCAAAAAA AAAAGCAAGA AGCGCAGTGG CTCACGCCTG TAATCCCAGC 12540

ACTTTGGGAG GCCGAGGCGG GCGGATCACG AGGTCAGGAG ATCGAGACCA TCCTGGCTAA 12600

CACGGTGAAA CCCCGTCTCT ACTAAAAATA CAAAAAATGA GCCGGGCGTG GTAGCGGGCG 12660

CCTGTAGTCC CAGCTACTCG GGAGGCTGAG GCAGGAGAAT GGCGTGAACC CGGGAGGCGG 12720

AGCTTGCAGT GAGCCGAGAT CGCGCCACTG CACTCCAGCC TGGGCGACAG AGCGAGACTC 12780

CGTNTCAAAA AAAAAAAAAA AAAAAAAAAC AAGAAAGAAA AAAAGAAGAT ACTGAAAAAT 12840

AGATGTCCCT AGTCAAAATA ATGAGATTAG CTTTTGACTA AACTCAGGAT ATTAAAAGGG 12900

AATACTTCAG TGCATGATGA TCTCATTTTT GAAAGGAAAG AANCAGAGCT TCCCCATCTC 12960

TAAAACCTTA ATTCAAAGGA GAAATAGATA ATTTCAAGAG GTATTTTTAT GAGGTAATAG 13020

TAAAATATAT TTTATTAACA GTACCTATAG TTATGTAAAA TAGGTAGTGC CAATTAACTG 13080

ACACTAAACT AGCTTCTTGG CCTGGCGCAG TGGCTCACGC CTGTNAATCC AAACACTTTG 13140

GGAGGCCGAT GCGGGTGTAT CGCTTGGGCT CAGGAATTCA AGGCCAGCCT GGGCAACATA 13200

TTAAAACCCC CTTTCTATAA AATATACAAA AATTAGCCAG GCATGGTGTG TGCCTGTAGT 13260

CCCAGATACT CAGGAGGCTG AGGCACGAGA ATCATGTGAA CCCAGGAGGT GGAGTTTGCA 13320

GTGAGCCGAG ATCACGCCAC TGCACTCCAG CCTGGGCAAC AGAGCAAAAC TCTGTCTCAA 13380

ATAATTAATA AATAAACTAG CTTCCTTTTC AAAAAAAGAA ATAAATTAGG TCCTAAGTCC 13440

TAAAAGCCCA TCCTACTTTA AAATTGTTTA TTCAAGTTCA GATGAAAAGA GTGGACTAGT 13500

AGGCAACTGA AGTGCTTTAG AGTCTCCCGT GCCTGCCCTA ATTTTAGAAG GTTGTGCACT 13560

TTATGATCCA GATTTCTGAG TGGTTGAGAA TGAGTTATTG AGCAGTGCAA GGCAAGCTCT 13620

GCAGTAGGTA ATGGATTGAT GAGGCTGGAT TTAGCAAGTC TGATCAATCT AAAGGAAGTT 13680

TCTGAATGTG TTTTTTGTAG TTAAAATACT CATAATTAAA ACACTTATCA CATTGTCACA 13740

TTTTATTTTT AAATTGCAGG TAAACAAGTG AGAACCAAAC TTTCACAGGC ATTTAATCAT 13800

TGGCTGAAAG TTCCAGAGGA CAAGCTACAG GTATTAGGCA ACTCTAACCT CATTAATCCC 13860

CAAGAAATTA ATAGCTGTCG CATAAAAATA TTCCTAGTTC TTGATTGAAT TTAGTCCTCA 13920

TGCAAGATAT TATTTTATAT TGAGGTTGCT AAATATTTAT TAGTTGTGAA AATTAACACA 13980

CCTGAGACTT TCATAATCTG TTAATTAAAC TGAGTAAGTT TTGAATAGTT CAAATAAGTG 14040

AAATTTTCAA TTTTTTTATT AGATTATTAT TGAAGTGACA GAAATGTTGC ATAATGCCAG 14100

TTTACTCATC GATGATATTG AAGACAACTC AAAACTCCGA CGTGGCTTTC CAGTGGCCCA 14160

CAGCATCTAT GGAATCCCAT CTGTCATCAA TTCTGCCAAT TACGTGTATT TCCTTGGCTT 14220

GGAGAAAGTC TTAACCCTTG ATCACCCAGA TGCAGTGAAG CTTTTTACCC GCCAGCTTTT 14280

GGAACTCCAT CAGGGACAAG GCCTAGATAT TTACTGGAGG GATAATTACA CTTGTCCCAC 14340

TGAAGAAGAA TATAAAGCTA TGGTGCTGCA GAAAACAGGT GGACTGTTTG GATTAGCAGT 14400

AGGTCTCATG CAGTTGTTCT CTGATTACAA AGAAGATTTA AAACCGCTAC TTAATACACT 14460

TGGGCTCTTT TTCCAAATTA GGGATGATTA TGCTAATCTA CACTCCAAAG AATATAGTGA 14520

AAACAAAAGT TTTTGTGAAG ATCTGACAGA GGGAAAGTTC TCATTTCCTA CTATTCATGC 14580

TATTTGGTCA AGGCCTGAAA GCACCCAGGT GCAGAATATC TTGCGCCAGA GAACAGAAAA 14640

CATAGATATA AAAAAATACT GTGTACATTA TCTTGAGGAT GTAGGTTCTT TTGAATACAC 14700

TCGTAATACC CTTAAAGAGC TTGAAGCTAA AGCCTATAAA CAGATTGATG CACGTGGTGG 14760

GAACCCTGAG CTAGTAGCCT TAGTAAAACA CTTAAGTAAG ATGTTCAAAG AAGAAAATGA 14820

ATAATGTTAA GCCATTCTTG ATTGGACCTC ATAGCTTATT TTAGTTAATC TTTNNTTTGT 14880

CTTTTAGCCT TACCACCTTT TAAAAAATTT GTTATTNTCC AGAAACAGTA AATAGGTGAG 14940

TAGGGGTGGT GCAAGTGAAT TCGTTTTCAT TTAGAAGCCC CTCTGTACAG ATAATCAAAA 15000

TTCAAAGTTG AAAGAATCAA AAGCAGCCAC AGTTATGTAG GTCTGATTTG AATGTCATAA 15060

TTGCAGTGAC AGGACATTGC CACCNNCTCG TATCCTACTA CCATCAATGT TGTGTTTATT 15120

CCGTCAATAA AAAAGACTTG CTTCCAGGAA TTTTTATCCA TACACTTTCT AACTGTACTA 15180

TCTGGGCAGT TCCAAGCCAG TTTCTATTAG CTAGCTGGAC CAAAGACCAC AAATCTCTTT 15240

TTTTCCTAAA CGCTGCTGTA AGGAATATCT CACTTTTCCC CCCGGAAACA CCCTCACTGA 15300

AGTCTTCTAT GAAAAGGCCT GATAATGGGC TGGGCGCGGT GGCTCACGCC TGTAATCCCA 15360

GCACTTTGGG AGGCCGAGGC GGGCAGATCA CGAGGTCAGG AGATCGAGAC CATCCTGACA 15420

CGGTGAAACC CTGTCTCTAC TAAAAATACA AAAAATTAGC TGGGCGTGGT GGTGGGCGCC 15480

TGTAGTCCCA GCTACTCGGG AGGCTGAGGC AGGAGAATGG TGTGAACCCA GGAGGCGGAG 15540

CTTGCAGTGA GCCGAGATAG TGCCTCTGCA CTCCAGCCTG GGTGACAGAG CGAGACTCCG 15600

TCTCAAAAAA AAGGGCTGAT AATGATAAAC AGTGAGCACT CCGGTCCTTT TTCTTAGGTT 15660

TTCCTTTTTT CCTTCCTCTC CACCCCACAA GTTTTGCTTT TTAACCAAGG TGTCTCTGCT 15720

TGATGAAATT CACATGCTAG TCTAAATCTT TTTTTCTCCC TTGTAACATT TATGTGCCCC 15780

AAACTGGTTA GTATATGGGT ACAGCATTCC CTTTCCAATT GGGAAGCGGA AAAAGAGAGT 15840

ATGGGATATT TTAGAAGGGA GCCTTTGAAC CTTATTATAT TTCCCCATCA TTGATAGTGA 15900

CAATCTTAAA AGGGTTGTTT TCTTACCTTA AGTACAAAAG CATGGAAAAA TGCGCTTTTC 15960

CTTCCCGCCC ACATCACCAC CCCGACTTGA AGACAGTAGG TGCTTGAATG GAAAGTGAGT 16020

AGGCATCTTT AATCGCCCTG ATTAAAGGAA AGTGTTAGCC TGAGAGGGCC TGACTGAAAA 16080

GTAACCAAAG GCTTAATATC AAACACTAAT TAGCTTTTTA GTGCCTTAAC CCTGACCTGG 16140

TTACCAGTTT TCTGTAGTTT CTACACCCAA GCCACTGAAG TCATCTGTGG CCCAAGAGGT 16200

AGGACAAAAA AAAAAAAAAA AAAAAAGCTG ATTTCAATAT TTGATTTGTT GACATCCCAA 16260

AATGAAAGTT TTATGTTTCC CTTAGAAACA TGTTTTGCTT GGTTCTATAG TATGTTACTT 16320

AGGATCTATT TACCATATAT TTGTATGAGA AATCCTCACC CAAGCATTCA ACCTAAATCT 16380

TTGAAAAGTT GGGTGCTGTC TTTAGTAACT TTTAAAATAG TTTAAATCTC CCATTTTAAT 16440

AGTGATAAGG AAACCTGTTA AAATCATGGC TATTGATGTT ATAGTATGGA AAGTTGAACT 16500

TTATGAACCC ATACTTTTAA AAAGCATTTT TAAAAATCTA ACACTGACTA TAGAAACAAA 16560

TTAAAATGTC TACCTTTAAG TATAAAAATT GCTTAAGTAG ATTTGTTCCT TGCCTATCAA 16620

ATTAATTTTG GCCTGGTGTT CTTCATTATT CATTTGTTAA TTTTATCTTG CCTTTGTCAA 16680

TAACAGAAAT GTTTGTCATT GAATTGGGAA TTTTTTTTTT TTTTTTTGAG ACGGAGTTTC 16740

ACTCTTGTTG CCCAGGCTGG AGTGCAATGG CGTGATCTCA GCTCACTGCA ACCTCCACCT 16800

CCCGGGTTCA AGCGATTCTC CTGCCTCAGC CTCCTAAGTA GCTGGGATTA CAGATGCCTG 16860

CCATGTTGCC TGGCTAATTT TTTTTTTTTT TTTTTTTTTA AGTAGAGATG GGGTTTCACC 16920

ATGTTGGCCA GGCTGGTGTT GAACTTCTGA CCTCAGGTGA TCCAGCTGCC TCGGCCTCCC 16980

AAAGTACTGG GATTACAGGC ATGAGCCACC GCACCCAGCC AAATTGGGGA CTTTTAACAG 17040

TCATTTTACC TGTAGAATAA TCAAAACTCT TCACTTGATC TGTAGTCATA GCTATTAACA 17100

CAGAAAAATG AATGCCAGTT ATGTTGCCAT A 17131

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 7314 base pairs
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: DOUBLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: DNA

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

TCGGGCTCCC TGGTTGGGGG GAGGGGGACG ACGAAAAATC CCCCCCGGAC TGGAGGTCCG 60

GGCCCCCAAT CGCGCTGCCC TCCAGAGGAC GGCGGCGATG GACCCTCTGC AGCTCCCTCC 120

GGGCAAAGGT CCAGGCGGTG GCCGTGGCGG CGGCAAGATG AAGCTCAAGA GTCTCCCTCC 180

GCTTCGGCGA CCGAGCTCCT CACTCCGGAC TCGACTGACG GGCAAACATC GCTTCCCCCC 240

CACCGACTCT AGGTTCCCCC CCTTTCTCCC CTCCCCTAGA TTTTTTTTCC CCCCCTCCCC 300

TACCTCTTTC CCGGATGGCC TCTTAGACGA CCTTGGATTG GTTAAAGTTC TTTAGAACCC 360

GCCTATACAC TGTTCCTATT GGTCCCTGGA TACAAACAAC GACGCCATTT TCCCACCAGT 420

TCTATGGAAA CAGAAAGTTA CGCCTCAAGG CTTTCTGGGA AATAAAGTCC ATACTCTGGG 480

GCCAACGCGC AAATCCTCGT CCGCGAGAAC TGCAAGGCCC GCAATGCCCT GCGCCTGCGT 540

GGACCGGTGC GGGGGCGGGG GGGAGGTGAA AGGGGCGGGG CAACAAAGCA GTAGGGAGGC 600

GGCAACGACG CCTGCGCAGT GTGACCGGGA TGGCGCATTT TCTTGCACCA ACTAATGCGG 660

TGTCGCTGGC GGCTGAGGAG GGCGGAGAGT TCTGTGGTGA AATAGTGGGA AGGATTCATG 720

TAGGCATCGG GAAGAGCCTA AGTCCACATT ATAAAATAGG AAGTTGATGC GGGGTACAGT 780

TACTCCCGGA CCGGCGGCGT GAAAGTCGTG ATATCATCGT TGAACTGTGA GCGGCAGTGG 840

CGGCGGCTGG GGGGAACCCG GATGGGAAGA AGGGCGGGGG AGGCTGGGAG GCGGGGCAGA 900

GGAAAGAAAG AAAGGAGAGT GAGGACCCGG ATGCTGAACC GGATTGTGTA TGAATTTTCC 960

ATCCCCTAGC TTTAAGCGAG GAGGGAGAGG AAGGGTTGGC CAAGTGGGGC GGAAGGGAGC 1020

ATCTGAGCGA GGAGGAAGCA GAAACCTCAC CGTTTCTTCC CCTCCGGACT CTGTGCTAGC 1080

ACTGTATACG TTTGCAGTTC TCTGCCCAGC CGCTGTGGAA AATCGGCCTC GAAGTGATTG 1140

AAATTCCCTG TTTATATCAG GCGGCTTCTT TCAGATCCAT CGTCTTTCTC CCGGAGTATG 1200

AATGGAAGGA TTCAGTATGC GCTTCACATT TGTATGTCTC TGGCCATTCT CAAACCAGGC 1260

CCTTCCCTTT GAAAAGTCTT TTGCATGGGA TGTTCACTTC TTAGACGCAA GGTTGTGTGC 1320

CCTGGTTTCA TCGTCTAACG CGTTAGAAGG CGCTTTCATT TCTTCATGGG TGTTGAGCGC 1380

CGACCACTGG GGTGGCCTCT GCCTTCGTAG ACCTGCGCCT GGTGAGACGG ACAGATGCTG 1440

AACAAAACGA TGTGAAATTA CCGCAGTGGC AGTGCCCCAG AGGAGAGTTC CACGGTGATA 1500

GGAGAATGAG GGAATTTGGC TTCTTTAGGG AGGGAAAGGA AGGGTTTCTG AGCAAGTGAG 1560

GATCGAGCTG AGAGCTGAAG GGCTAGCAGG AGTTAACTAA GGAAAGAGAA AAGGAAAAGA 1620

CATTCCAGAC AAAAAGGCTA ACTTGTCAGA AAGCCCTGTG GCGGAAGGGA GCTTTTCCAA 1680

TATGAAGAAC TGAGCCTGGA GAGATGGGAT GAGGGGGAGT GTCGAACCTT TTAGGCTTTG 1740

TAAAGGAGTT TTGGTTTTCT CCTAATAGCA ATGGGATATC TTCCAAGGAA TCTCAATCAA 1800

AAGGGAGAGA TGGCTCCGAT TGGAATGTCA TCCCTGGCTG AAGAGTNNAG GAAGCGAAAA 1860

AAAGAAGAGT TAAAGAGGCA AATGCAGGGA ACCCGACGAG GAGGCTATTG CCGTAGTAGT 1920

TCACATGGTG AAAAGAATGG AGCGTTTGTA TTAATGATTA TGGATTCACT CTTTGAACAA 1980

ATTTCTGGCA GCTTTTTAGT TTTGAAAGTG AGAAGTTTCA GACTCTCACT GAGGTATTCT 2040

GTAGTTTTTT CACTCTAAAA GGAAACTAGT AGAGTTCATG TAACACACAC TAATGCCTCT 2100

TTACATTTAA CTTTAGTATG TGATAGCTGA AATTTCCAGC TGTGATAAAT TGGGAAATCC 2160

TTTGATTTAA AAGAAAAACA AAGGCGGGTG AGGGTGAGAG TATATGCCAC GGTGTGTAGA 2220

ATCCTTTAGA CTCTTAAGAA GACACANGGC GGCTGGGCGT GGTGGCTCAC GCTTGTAATC 2280

CCAGCACTTT GGGAGGCCGA GGCGGGCGGA TCACGAGGTC AGGAGATCGA GACCATCCTG 2340

GCTAACACGG TGAAAGCCCG TCTCTACTAA AAATACAAAA AAATTAGCCG GGCAAGGTGG 2400

CGGGCGCCTG TAGTCCCAGC TACTCGGGAG GCTGAGGCAG GAGAATGGCG TGAACCCGGG 2460

AGGCGGAGTT TGCAGTGAGA CGAGATCACG CCACTGCACT CCAGCCTGGG CGACAGAGTG 2520

AGACGCTGTT TCAGAAGAAA GACACAAGGC AAGTTGGTTG TCGATACCTG GAAAAATTGA 2580

AGTTCTTATG TTTTCATACC ACTGAAAATG CTTGTATGTA AATATCCTCT GGGACAGGAA 2640

ATTGACTTAA GTGAGTATTC TTAAACATCT CTAAGTGAGG AAAGGAAATA TTTTTTAAAG 2700

CATAATTAGT GTTTTAAGTT GAAAAATAAC ATCAACCACA AAGCTCTACG AATTGAAACA 2760

AAGATTAGCT CTGATTTCTG TGCAACAGGG TACACCTGTT ACAGGTCCTG ACACAAAAGG 2820

GAATTCTGAA AGTGCATCTC ATTGATTTTT AAGTTCGGTC AAATGTGTTT TGGAGGCTGT 2880

GAGAAAATAT ACAAACGTGA TTCTTGCTCC CAACTTGTAG TTGAGAAAAG ATAGATACTA 2940

ACATTTAAAT AGAGAAGTAT ATGAGATCCT TTTTTAATTC TACTTTTAAT GATGTTCGAT 3000

AATAATCTTT TAGCTAAGCC ATTATTCTTC CTGTTTTGCA TCTTCTTTTC TTACTTCAAT 3060

CCCTGATAAT AAGGTCACGT GTCAGAGATC AAATAGTATA GGTAATAGGT TACCTAAATA 3120

GGTATTTGCA TAATAGGTTA CCTAACTAAA TAGGTTTTTG CCTAATAGGT ATGTTGATTA 3180

TTTCGCTTAC TTGATTCTTT ATGAGCCTTT TTTTCCTTGC GACGTCTTTG GTATTAATTG 3240

TTAGTCAAGA TGGATGTAGA AATTTTCCAT ATGGGATGTT TCTCTTTGAA TTCATGTTGT 3300

TAAAATGATT TCTTTTGGTG GAGTGCTGAT CTTTTTTATG ATTGTTTCAT ATAGATAAGA 3360

ACAGACTACA AAAAAATATG CCTTTCAATC CTGAAGAGTA ACCTGAACTA TACACTAGTT 3420

TTGTGCTTTA ATTTTCATTT GTAATCTGCC TTCAATAAAG AGTTAAGCTA GTGGAATTTA 3480

TGTCTTAGCT TGTTATAACA CAAACACGAA TATTTGTCTG CTTGGCATTA AAGGGTAAAG 3540

ATATTCCATA GCTGGGAATC TTAATCTGAG GTACGTGTAA ACATTCAGGG ACTATATGAT 3600

CTCTGAGAAT TTGTATGTTG TAAGTCTTTG TGGCAGTGTA TACATTTGTG TTGCAACTTA 3660

TTAACACATA CACCGGGCTT TTTTTTTTTT TTTTAGAAGA TTCATAGCTT TCATCATATT 3720

CTCAAAAGGT TTCTGTGACC CATGAGATGG TTTACAGTAT GGGGAAGCAT CAAAGCACTT 3780

GCACAGTTGA TGGTTATATG TGTGTGTTAT TATTTCAGCC ACCCATTATC ATGTGCTTAC 3840

CAACTGCCTA ACAGTGCATA CATATGTAGA AGTTTTATTC TTTTCTCCTG TTGCCATATT 3900

ATACGTCTCA TTTCACAGCA GAAAAACAAC TGCATGACAG AGACAATGTG GTTCAAACCA 3960

TTTTACCCTT GTATTCATTG ACTGCTACAA AACAGGAACA TTAAATACCT GATTGTCACC 4020

AAATTGGGTA GTCTCAGCAC TTCTACACTC GTAATTGTGC TGGAAAAGTG GAATGCTAGC 4080

ACTAATAATT AGATTTTGGT TTGGAGGGTT TTTTATTTGT TTATTCTTAC TTGTATAAAT 4140

TTATGGGGTG CAAGTGTAGT TTTATCACAT GCATAGATTG CATTGTAGTG AAGTCAGGAC 4200

TTTTAGGGGG TCCATCACCC ATGTAATCAC GTTGTACCCA TTAAGTAATC TTTCATCATC 4260

CACCTCCTTC CCACCTTCTC ACCCTTTGGA ATCTCCATTG TCTATCATTC CACACTCCAT 4320

GTCCATGTAT ACACATTATC TAGCTCCCAT TTATAATTGA GAAGATGTAC TATTTGTCTT 4380

TTATGTCTGA CTTGTTACAC TTAAGGTAAG GGCTATCCAT CCATTTTGCT GCAAATGACA 4440

TGATTTCATT TTGTTTTAAT GGCTGAGTAA TCATTCGTTG TATATATACC ACATTTTCTT 4500

TATTCAGTCA TCTGCTGATG GACACTTAGG TTGATTCCAT ATCTTTACTA TTGTGAATAG 4560

TGCTGTAATA AACACATAGT GCAAGATTTT GGAAATTTTA CTTTTGTGGC ACGTTGTTGG 4620

TATTTACTCA GGATCTTTGG ATTTGCTTGG CTGCATGTAT ATGAATCAGT GTGTTTATTT 4680

ACTGAAATAT GTGCAAAAGT CTTGTCTTTG GTGGATTAAT TTATAATATA AATCCACAAA 4740

AGTCAGATTC TGCTCCTAAG TATATTTTAC ATTTTTAAAT TTAATGCCAG CAAGAAGTTA 4800

CAGTACTAGA ATTGCCTTAC CCCTGAGAGT ATCAATGATC AGATCATAGT ATCAGGTGAC 4860

TGGGCTATAG AAGATGACTT TTATTACTTA ACATTATGAA GTTACTAGGG CTGATTTAGA 4920

AATCGAGGAA CACTGGTGAA ACCCCGTCTC TACTAAAATA CAAAAATTAG CTGGGCGTGG 4980

TGGTGGGCAC CTGTAGTCCC AGCTACTCAG AAGGCTGAGT CAGGAGAATT GCTTGAGCCC 5040

AGGAGGCAGA GGTTGCAGTG AGCCGAGATC GTGCCACTGC ACTCCAGCCT GGGCGACAGA 5100

GTGAGACTCC GTCTCAAAAA AAAAAAAAAA AAAAAAAAAG GAACACATCC TCACTGTTAC 5160

AATAAATAAC AGTAGCCCAC ACCCCCTTAG TTGTGATGTG GTGTGATACC ATGTAAGCAA 5220

CCTATTTCCA GTTCCCCTAA CATTCTCAAG CAGCTGTATC AGAATCATAC AAGATGCATA 5280

TTTAAATTGA AGATTTCTAA GTCTCTGGCC CAGACTTAGA AAAAAAGGAT CAGGCCGGGC 5340

ACAGTAGCTA ACACCTGCAA TTCCAACACT TTGGGAGGCT GAGGCGGGTG GATCGCCTGA 5400

GGTCAGGAGT TTTGAGACCA GCCTGGCCAA CATAGTGAAA CCCCATCTCT ACTAAAAATT 5460

CAAAAAATTA GCTGGGCGTG GTGGCAAGAA CCTGTAATCC CTGCTATTCG GGAGGCTGAG 5520

GCAGGGGAAT CACTTGAACC CGGGAGGTGG AGGTTGCAGT GAGCCAAGAT TGCGCCACTG 5580

CACTCCAGCC TGGGCAACGA GCAAAACTCC GTCTCAAAAA AAAAAAACAA AAGGACCTTT 5640

GAGCAATCAG AATAACACAA AGTACATGAA CTGAACTTCA TTTTCTTCAT TCAAAAGAAA 5700

GTGGCCCTCA CTCAAGCAAA TATATTCTTG TGCTTTATCT TCTGGCATAC TGAGATAACT 5760

TTCTAAAGTG GTTTCCAATT CCAAAATCCA ATGATGTGCA ACTCATTGAA CAGCCCTAAC 5820

CACAAACTGC CATTAGATGC CATATTACAT TTAGCCTTTT TGTTGTAGAA AAGTTGGTTA 5880

GAAGTGGGCT CAGGATTCTA AAGACTAAAT CATAGTCCCA AGAAGCAAAA GAAAGAGGAT 5940

AAAAGTAATA AACTTCCCAA AATGTGCCAA AGATGCTAGA GCAGTTAGAT TCCTAATATG 6000

AGGACAAGTA ATAATAGAAA CAGATACAAA GAAATAAAGT AGAGATTCAA CAGTACAGGG 6060

AGACCCTAGG AAGACCATGA GTGTTATTCT AGGAAATACT GAAATAAGAC AGATTTCAGT 6120

ATAAAGGGGN AATATGTTTA ATAANATATA TGCATTTGAG TTAATGCGTA TTTTAAATCA 6180

GAAATCTCTG AAATGGATTG ATTGTAGAGA AACTACTAGG GGGACGAGGA GAATCCCTTT 6240

AAATTTTAAA TACATAAAAC ATACTCATCT TAGTGCTCAT TTAAAAAAGG ATATGTTTAC 6300

TAATTAGTGT AATCAGTTAA ATACAGAGGT ATCTTTCCAA TTCTTTGGAT GTGTTTTGAC 6360

ATTTGCCGTC AACNAATTAA GCCTTTTGTG GTTGATTAAA ATAGGAAAAG CTTAATATAA 6420

GTTATGTGAC TAAGAAAACA ACTTAAAAAC CAAGACAACA CTTTGACCAA TATAATCACT 6480

TGAATGAAGA ATTTTCTAAT TGAGATATAA TTTACATACC ACCCATTTAA AGTGTACATT 6540

TCAGCAGTTT TTAGTGTATT CACAGGGCTG TGCAACCATC ACAATTTAAT TTTATAACAT 6600

TTTGATCCCT GCGAAAAGAA ACCCTGTACT CATTAGCAAT TAGTCCCTGT TCCTAACCAC 6660

TAATCTACTT TCTTTCTCTG TAGATTGGCT TATTCTGAAC ATTTCGTATA AATGGAATCA 6720

TACAATATGT AGTCTCTTGA GATTGGCTTC TTTCACTTAA CATGTTTTCA AGGCTTCATA 6780

GCTGTAGAAT CTTGCTTTGT TTTTTTGAGA CTGGAGTCAC TCTTTCGCCC AGGCTGGAGT 6840

GCAGTGGTGT GATCTCAGCT CACTGCAACC TCTGCCTCCC GGGTTCAAGC AGTTCTCCTG 6900

CCTCAGCCTC CCAAGTAGCC AGAACTACAG GCACACACCA CCATGCTCGG CTAATCTTTG 6960

TAGTTTTAGT AGAGATGGTG TGAAGGCTGG TCTCGAACTC CTGACCTCAT GATCTACCCA 7020

CCTCAGCTAA TTTTTCATAT TTTTAGTAGA GACAAGGTTT TGCCATGTTG CCCAGGCTGG 7080

TCTCGAACTC CTGGGCTTAA GCTATCCGCC CGCCTCAGCC TCCCAAAGTG CTGGGATTAC 7140

AGGCGTGAAC TACCGTGCCC AGCAACAGAA TCTTCTTTTT AAACCAGACT AGGTGTCTTT 7200

TCACAAACAC CCTGCAATAC AAATTCCTTT GCAGTTTGAC ACTGAAAGAT GATTAGTTTC 7260

ATGTGATCTT TATGTTTCTC CTTTTTGACA GATTAGCTTT GAAGTTTAAA TCCA 7314

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2307 base pairs
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: DOUBLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: DNA

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

TGTTAAGCCA TTCTTGATTG GACCTCATAG CTTATTTTAG TTAATCTTTN NTTTGTCTTT 60

TAGCCTTACC ACCTTTTAAA AAATTTGTTA TTNTCCAGAA ACAGTAAATA GGTGAGTAGG 120

GGTGGTGCAA GTGAATTCGT TTTCATTTAG AAGCCCCTCT GTACAGATAA TCAAAATTCA 180

AAGTTGAAAG AATCAAAAGC AGCCACAGTT ATGTAGGTCT GATTTGAATG TCATAATTGC 240

AGTGACAGGA CATTGCCACC NNCTCGTATC CTACTACCAT CAATGTTGTG TTTATTCCGT 300

CAATAAAAAA GACTTGCTTC CAGGAATTTT TATCCATACA CTTTCTAACT GTACTATCTG 360

GGCAGTTCCA AGCCAGTTTC TATTAGCTAG CTGGACCAAA GACCACAAAT CTCTTTTTTT 420

CCTAAACGCT GCTGTAAGGA ATATCTCACT TTTCCCCCCG GAAACACCCT CACTGAAGTC 480

TTCTATGAAA AGGCCTGATA ATGGGCTGGG CGCGGTGGCT CACGCCTGTA ATCCCAGCAC 540

TTTGGGAGGC CGAGGCGGGC AGATCACGAG GTCAGGAGAT CGAGACCATC CTGACACGGT 600

GAAACCCTGT CTCTACTAAA AATACAAAAA ATTAGCTGGG CGTGGTGGTG GGCGCCTGTA 660

GTCCCAGCTA CTCGGGAGGC TGAGGCAGGA GAATGGTGTG AACCCAGGAG GCGGAGCTTG 720

CAGTGAGCCG AGATAGTGCC TCTGCACTCC AGCCTGGGTG ACAGAGCGAG ACTCCGTCTC 780

AAAAAAAAGG GCTGATAATG ATAAACAGTG AGCACTCCGG TCCTTTTTCT TAGGTTTTCC 840

TTTTTTCCTT CCTCTCCACC CCACAAGTTT TGCTTTTTAA CCAAGGTGTC TCTGCTTGAT 900

GAAATTCACA TGCTAGTCTA AATCTTTTTT TCTCCCTTGT AACATTTATG TGCCCCAAAC 960

TGGTTAGTAT ATGGGTACAG CATTCCCTTT CCAATTGGGA AGCGGAAAAA GAGAGTATGG 1020

GATATTTTAG AAGGGAGCCT TTGAACCTTA TTATATTTCC CCATCATTGA TAGTGACAAT 1080

CTTAAAAGGG TTGTTTTCTT ACCTTAAGTA CAAAAGCATG GAAAAATGCG CTTTTCCTTC 1140

CCGCCCACAT CACCACCCCG ACTTGAAGAC AGTAGGTGCT TGAATGGAAA GTGAGTAGGC 1200

ATCTTTAATC GCCCTGATTA AAGGAAAGTG TTAGCCTGAG AGGGCCTGAC TGAAAAGTAA 1260

CCAAAGGCTT AATATCAAAC ACTAATTAGC TTTTTAGTGC CTTAACCCTG ACCTGGTTAC 1320

CAGTTTTCTG TAGTTTCTAC ACCCAAGCCA CTGAAGTCAT CTGTGGCCCA AGAGGTAGGA 1380

CAAAAAAAAA AAAAAAAAAA AAGCTGATTT CAATATTTGA TTTGTTGACA TCCCAAAATG 1440

AAAGTTTTAT GTTTCCCTTA GAAACATGTT TTGCTTGGTT CTATAGTATG TTACTTAGGA 1500

TCTATTTACC ATATATTTGT ATGAGAAATC CTCACCCAAG CATTCAACCT AAATCTTTGA 1560

AAAGTTGGGT GCTGTCTTTA GTAACTTTTA AAATAGTTTA AATCTCCCAT TTTAATAGTG 1620

ATAAGGAAAC CTGTTAAAAT CATGGCTATT GATGTTATAG TATGGAAAGT TGAACTTTAT 1680

GAACCCATAC TTTTAAAAAG CATTTTTAAA AATCTAACAC TGACTATAGA AACAAATTAA 1740

AATGTCTACC TTTAAGTATA AAAATTGCTT AAGTAGATTT GTTCCTTGCC TATCAAATTA 1800

ATTTTGGCCT GGTGTTCTTC ATTATTCATT TGTTAATTTT ATCTTGCCTT TGTCAATAAC 1860

AGAAATGTTT GTCATTGAAT TGGGAATTTT TTTTTTTTTT TTTGAGACGG AGTTTCACTC 1920

TTGTTGCCCA GGCTGGAGTG CAATGGCGTG ATCTCAGCTC ACTGCAACCT CCACCTCCCG 1980

GGTTCAAGCG ATTCTCCTGC CTCAGCCTCC TAAGTAGCTG GGATTACAGA TGCCTGCCAT 2040

GTTGCCTGGC TAATTTTTTT TTTTTTTTTT TTTTTAAGTA GAGATGGGGT TTCACCATGT 2100

TGGCCAGGCT GGTGTTGAAC TTCTGACCTC AGGTGATCCA GCTGCCTCGG CCTCCCAAAG 2160

TACTGGGATT ACAGGCATGA GCCACCGCAC CCAGCCAAAT TGGGGACTTT TAACAGTCAT 2220

TTTACCTGTA GAATAATCAA AACTCTTCAC TTGATCTGTA GTCATAGCTA TTAACACAGA 2280

AAAATGAATG CCAGTTATGT TGCCATA 2307

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1414 base pairs
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: cDNA

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

(ix) FEATURE:
(A) NAME/KEY: Polymorphic fragment 5-187-77 SEQ ID7
(B) LOCATION: 226..244

(ix) FEATURE:
(A) NAME/KEY: Polymorphic fragment 5-187-77 SEQ ID8
(B) LOCATION: 226..244

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:AA398854
(B) LOCATION: 1..477

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:AA435858
(B) LOCATION: complement 406..833

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:AA194600
(B) LOCATION: 1218..1414

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

```
CGCGCAAATC CTCGTCCGCG AGAACTGCAA GGCCCGCAAT GCCCTGCGCC TGCGTGGACC      60

GATTAGCTTT GAAGTTTAAA TCCA ATG GAG AAG ACT CAA GAA ACA GTC CAA          111
                           Met Glu Lys Thr Gln Glu Thr Val Gln
                           1               5

AGA ATT CTT CTA GAA CCC TAT AAA TAC TTA CTT CAG TTA CCA GGT AAA         159
Arg Ile Leu Leu Glu Pro Tyr Lys Tyr Leu Leu Gln Leu Pro Gly Lys
10                  15                  20                  25

CAA GTG AGA ACC AAA CTT TCA CAG GCA TTT AAT CAT TGG CTG AAA GTT         207
Gln Val Arg Thr Lys Leu Ser Gln Ala Phe Asn His Trp Leu Lys Val
                30                  35                  40

CCA GAG GAC AAG CTA CAG ATT ATT ATT GAA GTG ACA GAA ATG TTG CAT         255
Pro Glu Asp Lys Leu Gln Ile Ile Ile Glu Val Thr Glu Met Leu His
            45                  50                  55

AAT GCC AGT TTA CTC ATC GAT GAT ATT GAA GAC AAC TCA AAA CTC CGA         303
Asn Ala Ser Leu Leu Ile Asp Asp Ile Glu Asp Asn Ser Lys Leu Arg
        60                  65                  70

CGT GGC TTT CCA GTG GCC CAC AGC ATC TAT GGA ATC CCA TCT GTC ATC         351
Arg Gly Phe Pro Val Ala His Ser Ile Tyr Gly Ile Pro Ser Val Ile
    75                  80                  85

AAT TCT GCC AAT TAC GTG TAT TTC CTT GGC TTG GAG AAA GTC TTA ACC         399
Asn Ser Ala Asn Tyr Val Tyr Phe Leu Gly Leu Glu Lys Val Leu Thr
90                  95                  100                 105

CTT GAT CAC CCA GAT GCA GTG AAG CTT TTT ACC CGC CAG CTT TTG GAA         447
Leu Asp His Pro Asp Ala Val Lys Leu Phe Thr Arg Gln Leu Leu Glu
                110                 115                 120

CTC CAT CAG GGA CAA GGC CTA GAT ATT TAC TGG AGG GAT AAT TAC ACT         495
Leu His Gln Gly Gln Gly Leu Asp Ile Tyr Trp Arg Asp Asn Tyr Thr
            125                 130                 135
```

```
TGT CCC ACT GAA GAA GAA TAT AAA GCT ATG GTG CTG CAG AAA ACA GGT      543
Cys Pro Thr Glu Glu Glu Tyr Lys Ala Met Val Leu Gln Lys Thr Gly
        140                 145                 150

GGA CTG TTT GGA TTA GCA GTA GGT CTC ATG CAG TTG TTC TCT GAT TAC      591
Gly Leu Phe Gly Leu Ala Val Gly Leu Met Gln Leu Phe Ser Asp Tyr
    155                 160                 165

AAA GAA GAT TTA AAA CCG CTA CTT AAT ACA CTT GGG CTC TTT TTC CAA      639
Lys Glu Asp Leu Lys Pro Leu Leu Asn Thr Leu Gly Leu Phe Phe Gln
170                 175                 180                 185

ATT AGG GAT GAT TAT GCT AAT CTA CAC TCC AAA GAA TAT AGT GAA AAC      687
Ile Arg Asp Asp Tyr Ala Asn Leu His Ser Lys Glu Tyr Ser Glu Asn
                190                 195                 200

AAA AGT TTT TGT GAA GAT CTG ACA GAG GGA AAG TTC TCA TTT CCT ACT      735
Lys Ser Phe Cys Glu Asp Leu Thr Glu Gly Lys Phe Ser Phe Pro Thr
            205                 210                 215

ATT CAT GCT ATT TGG TCA AGG CCT GAA AGC ACC CAG GTG CAG AAT ATC      783
Ile His Ala Ile Trp Ser Arg Pro Glu Ser Thr Gln Val Gln Asn Ile
        220                 225                 230

TTG CGC CAG AGA ACA GAA AAC ATA GAT ATA AAA AAA TAC TGT GTA CAT      831
Leu Arg Gln Arg Thr Glu Asn Ile Asp Ile Lys Lys Tyr Cys Val His
    235                 240                 245

TAT CTT GAG GAT GTA GGT TCT TTT GAA TAC ACT CGT AAT ACC CTT AAA      879
Tyr Leu Glu Asp Val Gly Ser Phe Glu Tyr Thr Arg Asn Thr Leu Lys
250                 255                 260                 265

GAG CTT GAA GCT AAA GCC TAT AAA CAG ATT GAT GCA CGT GGT GGG AAC      927
Glu Leu Glu Ala Lys Ala Tyr Lys Gln Ile Asp Ala Arg Gly Gly Asn
                270                 275                 280

CCT GAG CTA GTA GCC TTA GTA AAA CAC TTA AGT AAG ATG TTC AAA GAA      975
Pro Glu Leu Val Ala Leu Val Lys His Leu Ser Lys Met Phe Lys Glu
            285                 290                 295

GAA AAT GAA TAA TGTTAAGCCA TTCTTGATTG GACCTCATAG CTTATTTTAG         1027
Glu Asn Glu  *
        300

TTAATCTTTN NTTTGTCTTT TAGCCTTACC ACCTTTTAAA AAATTTGTTA TTNTCCAGAA   1087

ACAGTAAATA GGTGAGTAGG GGTGGTGCAA GTGAATTCGT TTTCATTTAG AAGCCCCTCT   1147

GTACAGATAA TCAAAATTCA AAGTTGAAAG AATCAAAAGC AGCCACAGTT ATGTAGGTCT   1207
```

GATTTGAATG TCATAATTGC AGTGACAGGA CATTGCCACC NNCTCGTATC CTACTACCAT 1267

CAATGTTGTG TTTATTCCGT CAATAAAAAA GACTTGCTTC CAGGAATTTT TATCCATACA 1327

CTTTCTAACT GTACTATCTG GGCAGTTCCA AGCCAGTTTC TATTAGCTAG CTGGACCAAA 1387

GACCACAAAT CTCTTTTTTT CCTAAAC 1414

(2) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1547 base pairs
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: cDNA

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

(ix) FEATURE:
(A) NAME/KEY: homology with EST in ref embl:Z44596
(B) LOCATION: 1..359

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

GCGCATTTTC TTGCACCAAC TAATGCGGTG TCGCTGGCGG CTGAGGAGGG CGGAGAGTTC 60

TGTGGTGAAA TAGTGGGAAG GATTCATGTA GGCATCGGGA AGAGCCTAAG TCCACATTAT 120

AAAATAGGAA GTTGATGCGG GGTACAGTTA CTCCCGGACC GGCGGCGTGA AAGTCGTGAT 180

ATCATCGTTG AACTATTAGC TTTGAAGTTT AAATCCA ATG GAG AAG ACT CAA GAA 235
Met Glu Lys Thr Gln Glu
1 5

ACA GTC CAA AGA ATT CTT CTA GAA CCC TAT AAA TAC TTA CTT CAG TTA 283
Thr Val Gln Arg Ile Leu Leu Glu Pro Tyr Lys Tyr Leu Leu Gln Leu
10 15 20

CCA GGT AAA CAA GTG AGA ACC AAA CTT TCA CAG GCA TTT AAT CAT TGG 331
Pro Gly Lys Gln Val Arg Thr Lys Leu Ser Gln Ala Phe Asn His Trp
25 30 35

```
CTG AAA GTT CCA GAG GAC AAG CTA CAG ATT ATT ATT GAA GTG ACA GAA      379
Leu Lys Val Pro Glu Asp Lys Leu Gln Ile Ile Ile Glu Val Thr Glu
    40                  45                  50

ATG TTG CAT AAT GCC AGT TTA CTC ATC GAT GAT ATT GAA GAC AAC TCA      427
Met Leu His Asn Ala Ser Leu Leu Ile Asp Asp Ile Glu Asp Asn Ser
55                  60                  65                  70

AAA CTC CGA CGT GGC TTT CCA GTG GCC CAC AGC ATC TAT GGA ATC CCA      475
Lys Leu Arg Arg Gly Phe Pro Val Ala His Ser Ile Tyr Gly Ile Pro
                75                  80                  85

TCT GTC ATC AAT TCT GCC AAT TAC GTG TAT TTC CTT GGC TTG GAG AAA      523
Ser Val Ile Asn Ser Ala Asn Tyr Val Tyr Phe Leu Gly Leu Glu Lys
            90                  95                  100

GTC TTA ACC CTT GAT CAC CCA GAT GCA GTG AAG CTT TTT ACC CGC CAG      571
Val Leu Thr Leu Asp His Pro Asp Ala Val Lys Leu Phe Thr Arg Gln
        105                 110                 115

CTT TTG GAA CTC CAT CAG GGA CAA GGC CTA GAT ATT TAC TGG AGG GAT      619
Leu Leu Glu Leu His Gln Gly Gln Gly Leu Asp Ile Tyr Trp Arg Asp
    120                 125                 130

AAT TAC ACT TGT CCC ACT GAA GAA GAA TAT AAA GCT ATG GTG CTG CAG      667
Asn Tyr Thr Cys Pro Thr Glu Glu Glu Tyr Lys Ala Met Val Leu Gln
135                 140                 145                 150

AAA ACA GGT GGA CTG TTT GGA TTA GCA GTA GGT CTC ATG CAG TTG TTC      715
Lys Thr Gly Gly Leu Phe Gly Leu Ala Val Gly Leu Met Gln Leu Phe
                155                 160                 165

TCT GAT TAC AAA GAA GAT TTA AAA CCG CTA CTT AAT ACA CTT GGG CTC      763
Ser Asp Tyr Lys Glu Asp Leu Lys Pro Leu Leu Asn Thr Leu Gly Leu
            170                 175                 180

TTT TTC CAA ATT AGG GAT GAT TAT GCT AAT CTA CAC TCC AAA GAA TAT      811
Phe Phe Gln Ile Arg Asp Asp Tyr Ala Asn Leu His Ser Lys Glu Tyr
        185                 190                 195

AGT GAA AAC AAA AGT TTT TGT GAA GAT CTG ACA GAG GGA AAG TTC TCA      859
Ser Glu Asn Lys Ser Phe Cys Glu Asp Leu Thr Glu Gly Lys Phe Ser
    200                 205                 210

TTT CCT ACT ATT CAT GCT ATT TGG TCA AGG CCT GAA AGC ACC CAG GTG      907
Phe Pro Thr Ile His Ala Ile Trp Ser Arg Pro Glu Ser Thr Gln Val
215                 220                 225                 230
```

```
CAG AAT ATC TTG CGC CAG AGA ACA GAA AAC ATA GAT ATA AAA AAA TAC      955
Gln Asn Ile Leu Arg Gln Arg Thr Glu Asn Ile Asp Ile Lys Lys Tyr
                235                 240                 245

TGT GTA CAT TAT CTT GAG GAT GTA GGT TCT TTT GAA TAC ACT CGT AAT     1003
Cys Val His Tyr Leu Glu Asp Val Gly Ser Phe Glu Tyr Thr Arg Asn
            250                 255                 260

ACC CTT AAA GAG CTT GAA GCT AAA GCC TAT AAA CAG ATT GAT GCA CGT     1051
Thr Leu Lys Glu Leu Glu Ala Lys Ala Tyr Lys Gln Ile Asp Ala Arg
        265                 270                 275

GGT GGG AAC CCT GAG CTA GTA GCC TTA GTA AAA CAC TTA AGT AAG ATG     1099
Gly Gly Asn Pro Glu Leu Val Ala Leu Val Lys His Leu Ser Lys Met
    280                 285                 290

TTC AAA GAA GAA AAT GAA TAA TGTTAAGCCA TTCTTGATTG GACCTCATAG        1150
Phe Lys Glu Glu Asn Glu  *
295                 300

CTTATTTTAG TTAATCTTTN NTTTGTCTTT TAGCCTTACC ACCTTTTAAA AAATTTGTTA   1210

TTNTCCAGAA ACAGTAAATA GGTGAGTAGG GGTGGTGCAA GTGAATTCGT TTTCATTTAG   1270

AAGCCCCTCT GTACAGATAA TCAAAATTCA AAGTTGAAAG AATCAAAAGC AGCCACAGTT   1330

ATGTAGGTCT GATTTGAATG TCATAATTGC AGTGACAGGA CATTGCCACC NNCTCGTATC   1390

CTACTACCAT CAATGTTGTG TTTATTCCGT CAATAAAAAA GACTTGCTTC CAGGAATTTT   1450

TATCCATACA CTTTCTAACT GTACTATCTG GGCAGTTCCA AGCCAGTTTC TATTAGCTAG   1510

CTGGACCAAA GACCACAAAT CTCTTTTTTT CCTAAAC                            1547
```

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 300 amino acids
- (B) TYPE: AMINO ACID
- (C) STRANDEDNESS: SINGLE
- (D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: protein

(vi) ORIGINAL SOURCE:
- (A) ORGANISM: Homo sapiens

(ix) FEATURE:
(A) NAME/KEY: diverging amino acid, Leu in ref genseqp:R97565
(B) LOCATION: 204

(ix) FEATURE:
(A) NAME/KEY: diverging amino acid, Gly in ref genseqp:R97565
(B) LOCATION: 205

(ix) FEATURE:
(A) NAME/KEY: diverging amino acid, Ser in ref genseqp:R97565
(B) LOCATION: 225

(ix) FEATURE:
(A) NAME/KEY: diverging amino acid, Lys in ref genseqp:R97565
(B) LOCATION: 252

(ix) FEATURE:
(A) NAME/KEY: diverging amino acid, Gly in ref genseqp:R97565
(B) LOCATION: 257

(ix) FEATURE:
(A) NAME/KEY: diverging amino acid, Ser in ref genseqp:R97565
(B) LOCATION: 295

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

```
Met Glu Lys Thr Gln Glu Thr Val Gln Arg Ile Leu Leu Glu Pro Tyr
1               5                  10                  15

Lys Tyr Leu Leu Gln Leu Pro Gly Lys Gln Val Arg Thr Lys Leu Ser
            20                  25                  30

Gln Ala Phe Asn His Trp Leu Lys Val Pro Glu Asp Lys Leu Gln Ile
        35                  40                  45

Ile Ile Glu Val Thr Glu Met Leu His Asn Ala Ser Leu Leu Ile Asp
    50                  55                  60

Asp Ile Glu Asp Asn Ser Lys Leu Arg Arg Gly Phe Pro Val Ala His
65                  70                  75                  80

Ser Ile Tyr Gly Ile Pro Ser Val Ile Asn Ser Ala Asn Tyr Val Tyr
                85                  90                  95

Phe Leu Gly Leu Glu Lys Val Leu Thr Leu Asp His Pro Asp Ala Val
            100                 105                 110
```

Lys Leu Phe Thr Arg Gln Leu Leu Glu Leu His Gln Gly Gln Gly Leu
115 120 125

Asp Ile Tyr Trp Arg Asp Asn Tyr Thr Cys Pro Thr Glu Glu Glu Tyr
130 135 140

Lys Ala Met Val Leu Gln Lys Thr Gly Gly Leu Phe Gly Leu Ala Val
145 150 155 160

Gly Leu Met Gln Leu Phe Ser Asp Tyr Lys Glu Asp Leu Lys Pro Leu
165 170 175

Leu Asn Thr Leu Gly Leu Phe Phe Gln Ile Arg Asp Asp Tyr Ala Asn
180 185 190

Leu His Ser Lys Glu Tyr Ser Glu Asn Lys Ser Phe Cys Glu Asp Leu
195 200 205

Thr Glu Gly Lys Phe Ser Phe Pro Thr Ile His Ala Ile Trp Ser Arg
210 215 220

Pro Glu Ser Thr Gln Val Gln Asn Ile Leu Arg Gln Arg Thr Glu Asn
225 230 235 240

Ile Asp Ile Lys Lys Tyr Cys Val His Tyr Leu Glu Asp Val Gly Ser
245 250 255

Phe Glu Tyr Thr Arg Asn Thr Leu Lys Glu Leu Glu Ala Lys Ala Tyr
260 265 270

Lys Gln Ile Asp Ala Arg Gly Gly Asn Pro Glu Leu Val Ala Leu Val
275 280 285

Lys His Leu Ser Lys Met Phe Lys Glu Glu Asn Glu
290 295 300

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 49 base pairs
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: DNA

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

(ix) FEATURE:
(A) NAME/KEY: polymorphic fragment 5-187-77
(B) LOCATION: 1..49

(ix) FEATURE:
(A) NAME/KEY: polymorphic base
(B) LOCATION: 23
(D) OTHER INFORMATION: insertion of a T in SEQID8

(ix) FEATURE:
(A) NAME/KEY: microsequencing oligo 5-187-77
(B) LOCATION: 4..22

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

AAGTGAAATT TTCAATTTTT TTATTAGATT ATTATTGAAG TGACAGAAA 49

(2) INFORMATION FOR SEQ ID NO: 8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 50 base pairs
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: DNA

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

(ix) FEATURE:
(A) NAME/KEY: polymorphic fragment 5-187-77
(B) LOCATION: 1..50
(D) OTHER INFORMATION: variant version of SEQ ID7

(ix) FEATURE:
(A) NAME/KEY: polymorphic base
(B) LOCATION: 23
(D) OTHER INFORMATION: base inerted T

(ix) FEATURE:
(A) NAME/KEY: Potential microsequencing oligo 5-187-77
(B) LOCATION: 4..22

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:

AAGTGAAATT TTCAATTTTT TTTATTAGAT TATTATTGAA GTGACAGAAA 50

(2) INFORMATION FOR SEQ ID NO: 9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: DNA

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

(ix) FEATURE:
(A) NAME/KEY: upstream amplification primer for SEQ ID7, SEQ ID8
(B) LOCATION: 1..19

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 9:

CTGAGACTTT CATAATCTG 19

(2) INFORMATION FOR SEQ ID NO: 10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: DNA

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

(ix) FEATURE:
(A) NAME/KEY: downstream amplification primer for SEQ ID7, SEQ ID8
(B) LOCATION: 1..20

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 10:

ATGAGACCTA CTGCTAATCC 20

(2) INFORMATION FOR SEQ ID NO: 11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: NUCLEIC ACID
(C) STRANDEDNESS: SINGLE
(D) TOPOLOGY: LINEAR

(ii) MOLECULE TYPE: DNA

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

(ix) FEATURE:
(A) NAME/KEY: microsequencing oligo 5-187-77.mis1
(B) LOCATION: 1..19

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 11:

TGAAATTTTC AATTTTTTT 19

FIGURE 1

GGPPS genome structure and cDNAs

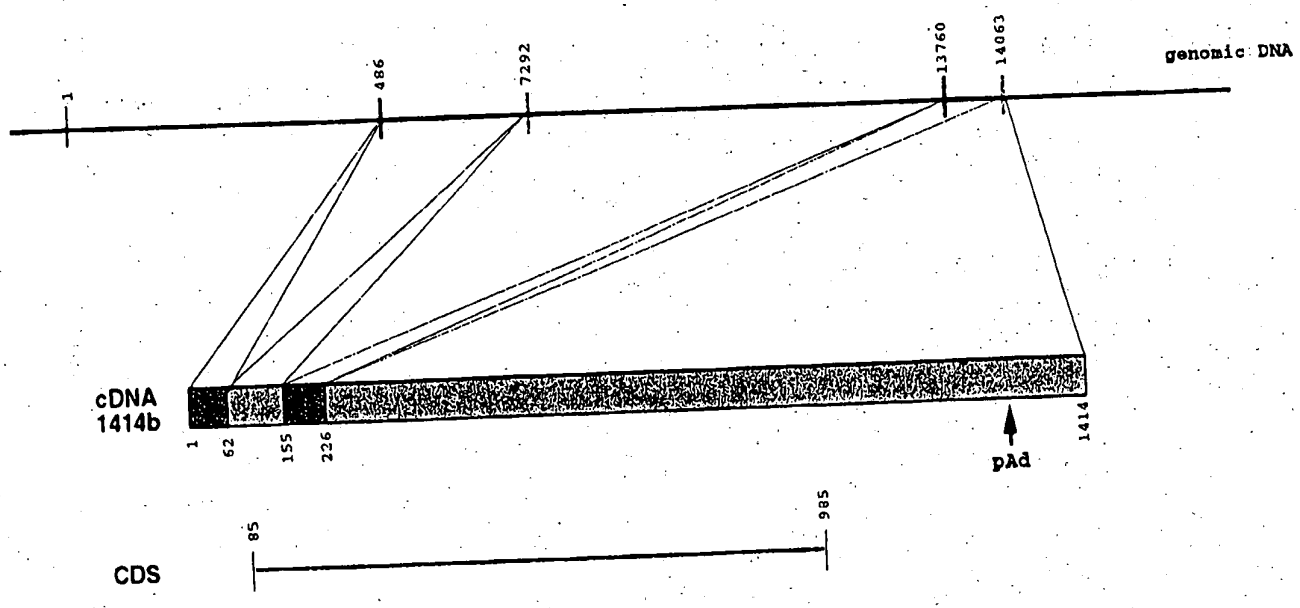

FIGURE 2

GGPPS genome structure and Variant cDNA